Consistent Validation for Predictive Methods in Spatial Settings

David R. Burt [1] **Yunyi Shen** [1] **Tamara Broderick** [1]

Abstract

Spatial prediction tasks are key to weather forecasting, studying air pollution impacts, and other scientific endeavors. Determining how much to trust predictions made by statistical or physical methods is essential for the credibility of scientific conclusions. Unfortunately, classical approaches for validation fail to handle mismatch between locations available for validation and (test) locations where we want to make predictions. This mismatch is often not an instance of covariate shift (as commonly formalized) because the validation and test locations are fixed (e.g., on a grid or at select points) rather than i.i.d. from two distributions. In the present work, we formalize a check on validation methods: that they become arbitrarily accurate as validation data becomes arbitrarily dense. We show that classical and covariate-shift methods can fail this check. We instead propose a method that builds from existing ideas in the covariate-shift literature, but adapts them to the validation data at hand. We prove that our proposal passes our check. And we demonstrate its advantages empirically on simulated and real data.

1. Introduction

Researchers are often interested in making predictions in a spatial setting. For instance, scientists predict sea surface temperature (SST) for weather forecasting and climate research (Minnett, 2010), predict air pollution at population centers to better understand the effect of pollution on health outcomes such as kidney disease (Remigio et al., 2022), or predict the prevalence of an invasive species for ecological management (Barbet-Massin et al., 2018). Characterizing the reliability of these predictions is key to understanding their suitability for downstream applications; e.g., Minnett (2010) describes acceptable SST error tolerances for weather forecasting. Estimates of prediction accuracy can also be used to choose between several predictive methods, as in Shabani et al. (2016).

In the spatial setting, predictive methods need not always arise from a statistical or machine learning approach built using training data. The predictive method is often a complex physical model provided by a third party (Remigio et al., 2022; Minnett, 2010; Gupta et al., 2018). Or it could combine physics and data-driven models (Banzon et al., 2016; Werner et al., 2019; Özkaynak et al., 2013).

In any of these cases, it is common to estimate the performance of a predictive method by using a set of *validation data*. More precisely, we are ultimately interested in predicting a response at what we call *test sites*; in the SST example above (Minnett, 2010), the test sites are points on a grid (often called a map), or in the air pollution example (Remigio et al., 2022), the test sites are 28 counties in the US Northeast. We can make predictions at the test sites, but we do not have access to direct observations of the responses there. We do have observed responses in the validation data, which we assume were not used in forming the predictive method being evaluated; in the SST example, scientists have SST observations taken by boats and buoys as validation data. So our aim is to estimate the average loss (i.e. risk) at the test sites using the validation data. We will see that many popular or natural approaches fail at this task.

One widely used approach, called the *holdout*,[1] estimates the test risk by taking the empirical average of the validation loss. When the validation and test data are independent and identically distributed (i.i.d.) from the same distribution, the holdout has a rigorous justification (Devroye, 1976; Langford, 2005). But in spatial problems, the validation and test sites need not be similarly dispersed, all data may be spatially correlated, and the test sites are often fixed rather than random; recall the grid or point prediction examples above. Indeed, Roberts et al. (2017) have observed problems with the holdout in practice. In the special case where the

[1]Laboratory for Information and Decision Systems, Massachusetts Institute for Technology, Cambridge MA, USA. Correspondence to: David R. Burt <dburt@mit.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]The name originally referred to "holding out" data for validation, with the remainder of available data going toward training. While we maintain the naming convention, we emphasize that in our setup there need not be any training data. Minnett (2010); Gupta et al. (2018); Duan et al. (2019), among many others, use this approach.

predictive method is data-driven, some authors (e.g. Telford & Birks, 2005) have suggested choosing holdout validation sites far from the training data sites. But we expect this proposal to still suffer from the problems just mentioned, and in fact simulation studies suggest it can misestimate test risk (Ploton et al., 2020; De Bruin et al., 2022).

Another natural idea is to use covariate-shift approaches to handle potential mismatch between validation and test sites (Sarafian et al., 2020; De Bruin et al., 2022). However, the covariate-shift literature generally assumes validation sites are drawn i.i.d. from one distribution, test sites are i.i.d. from another, and the density ratio between these two distributions exists and is bounded. For both the grid and point examples, these last two assumptions are inappropriate.

In what follows, we lay out a precise formulation of the prediction validation task in the spatial setting (Section 2). We formalize a desirable property for test-risk estimators: that, if arbitrarily dense validation data accrues in a region including the test points, the test-risk estimate should become arbitrarily accurate (Section 3). We prove that both the holdout estimator and an estimator advocated in the covariate-shift literature (Loog, 2012; Portier et al., 2023) fail to satisfy this spatial consistency property (Section 4). We propose to build on the k-nearest neighbor estimator (Loog, 2012). In particular, Loog (2012) and Portier et al. (2023) advocated fixing $k = 1$ for covariate-shift problems. We instead derive an upper bound on the error of the general-k estimator for estimating test risk (Section 5.1); crucially, our bound is *conditional on the test and validation sites*. We prove that choosing k adaptively by optimizing our upper bound yields a spatially consistent estimator (Section 5.2). Unlike covariate-shift results (e.g. Portier et al., 2023), our results are directly applicable to problems where the test sites are most reasonably thought of as fixed. We illustrate the accuracy and practicality of our proposed method in simulated and real data analyses (Section 6), with tasks in both grid and point prediction. We discuss further related work in App. B.

2. Estimating Test Risk in a Spatial Problem

We now formalize risk estimation at test points in a spatial setting. We assume each data point occurs at a spatial location $S \in \mathcal{S}$, where the spatial domain $(\mathcal{S}, d_{\mathcal{S}})$ is a metric space. Each data point has observed covariates $X \in \mathcal{X}$ and a response $Y \in \mathcal{Y}$. The covariates are a fixed spatial field, $\chi : \mathcal{S} \to \mathcal{X}$; i.e., the covariates at a point are specified by evaluating χ at the point's spatial location.

2.1. Test risk of a spatial predictive method

We assume we have access to a predictive method $h : (\mathcal{S}, \mathcal{X}) \to \mathcal{Y}$. We define $h^{\chi} : \mathcal{S} \to \mathcal{Y}$ to be the prediction made by h at the location S: $h^{\chi}(S) = h(S, \chi(S))$. We suppose that practitioners would like to use h to predict the response at a set of test sites where the response is unknown. We collect the M^{test} test data points, including true (but unobserved) responses, in $D^{\text{test}} = (S_m^{\text{test}}, X_m^{\text{test}}, Y_m^{\text{test}})_{m=1}^{M^{\text{test}}}$.

To quantify quality of a predictive method, we need a loss. We assume the loss is bounded, as is often the case for practically-bounded responses; cf. temperature, pressure, or other physical quantities.[2]

Assumption 2.1 (Δ-bounded Loss)**.** The loss is a non-negative, bounded function, $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \Delta]$.

Due to practical considerations such as measurement error, the response at a test point is usefully modeled as random. To summarize loss over this randomness, it is standard to consider expected loss (a.k.a. risk) at the testing data. To define this expectation, we need to make assumptions about the data-generating process. It is typical in the non-spatial setting to assume responses are i.i.d. conditional on covariates. In the spatial setting, the i.i.d. assumption is inappropriate since it ignores spatial location. We instead assume that the response variable may be a function of the spatial location it is observed at, the covariates at that location, and i.i.d noise:

Assumption 2.2 (Data Generating Process: Test Data)**.** Let $j = \text{test}$. Let $\chi : \mathcal{S} \to \mathcal{X}$ be a fixed function. For $1 \leq m \leq M^j$, $X_m^j = \chi(S_m^j)$ and $Y_m^j = f(S_m^j, X_m^j, \epsilon_m^j)$ with $f : \mathcal{S} \times \mathcal{X} \to \mathcal{Y}$ and $\epsilon_m^j \overset{\text{iid}}{\sim} P_\epsilon$ real-valued random variables.

Assumption 2.2 implies the response is i.i.d. *given* the spatial location. For example, Assumption 2.2 would cover the case where measurement errors on sensors are independent, but the locations of the sensors are not. A widely studied special case of Assumption 2.2 considers additive, homoskedastic noise: namely, $\mathcal{Y} \subset \mathbb{R}$ and $Y_m^j = f(S_m^j, X_m^j) + \epsilon_m^j$. Assumption 2.2 is more general; for example, it allows the noise to be scaled by a continuous, deterministic function of $\mathcal{S}$: $Y_m^j = f(S_m^j, X_m^j) + g(S_m^j)\epsilon_m^j$. Before defining risk, we first define the average loss of the predictive method at a particular location in space, S: $e_h(S) := \mathbb{E}[\ell(f(S, \chi(S), \epsilon), h^{\chi}(S))|S]$. Finally, we average over all spatial locations of interest, which we assume is a finite set.

Definition 2.3. Given test points $(S_m^{\text{test}})_{m=1}^{M^{\text{test}}}$, let $Q^{\text{test}} := (1/M^{\text{test}}) \sum_{m=1}^{M^{\text{test}}} \delta_{S_m^{\text{test}}}$, with δ_S a Dirac measure at S. For predictive method h, let the *test risk* of h be $R_{Q^{\text{test}}}(h) := (1/M^{\text{test}}) \sum_{m=1}^{M^{\text{test}}} e_h(S_m^{\text{test}})$.

[2] Loss is also bounded for classification error or robust regression cases such as Tukey's biweight loss.

2.2. Estimating test risk

To estimate test risk, we assume we have access to N^{val} validation data points, collected in $D^{\text{val}} = (S_n^{\text{val}}, X_n^{\text{val}}, Y_n^{\text{val}})_{n=1}^{N^{\text{val}}}$. We assume practitioners did not make use of either the validation or test response data when constructing the predictive method. As an example, the common holdout estimator uses the empirical average of validation loss:

$$\hat{R}_{\text{Hold}}(h) \coloneqq (1/N^{\text{val}}) \sum_{n=1}^{N^{\text{val}}} \ell(Y_n^{\text{val}}, h^{\chi}(S_n^{\text{val}})). \quad (1)$$

For validation data to provide information about test risk, we need regularity assumptions. First, we make a standard assumption that validation data follows the same data-generating process as test data. And second, we assume a form of smoothness across the spatial locations.

Assumption 2.4 (Data Generating Process: Validation Data)**.** Assumption 2.2 remains true when we take $j = \text{val}$, with the same f, χ, and P_ϵ as for $j = \text{test}$.

Assumption 2.5 (L-Lipschitz)**.** For some $L \geq 0$, for all $S, S' \in \mathcal{S}, |e_h(S) - e_h(S')| \leq L d_{\mathcal{S}}(S, S')$.

Assumption 2.5 often arises naturally. For example, consider $\mathcal{Y} \subset [0,1]$, $\mathcal{S} = (\mathbb{R}^d, \|\cdot\|_2)$, squared loss, and homoskedastic and additive noise. Suppose $f(S, \chi(S))$ is L_Y-Lipschitz and h^{χ} is L_h-Lipschitz. Then Assumption 2.5 holds with $L = 2(L_Y + L_h)$; see Prop. D.1.

3. We Want Consistent Estimators

Once we have an estimator of test risk, it remains to check if that estimator performs well. We next formalize one natural check on performance: namely, estimators should become arbitrarily accurate if given validation data that is arbitrarily dense in the spatial domain. This check is analogous to traditional consistency in the i.i.d. data setting.

To that end, note that the *fill distance* is a measure of discrepancy between two sets of points, Ψ_1 and Ψ_2.[3] It is the maximum distance from a point in Ψ_2 to the nearest point in Ψ_1.

Definition 3.1 (Cressie 2015, §5.8, Wendland 2004, Definition 1.4)**.** Let $(\mathcal{S}, d_{\mathcal{S}})$ be a metric space and $\Psi_1, \Psi_2 \subset \mathcal{S}$. The fill distance of Ψ_1 in Ψ_2 is $\zeta(\Psi_1; \Psi_2) := \sup_{S_2 \in \Psi_2} \inf_{S_1 \in \Psi_1} d_{\mathcal{S}}(S_1, S_2)$.

In the spatial statistics literature, *infill asymptotics* describes cases where data are gathered over a compact spatial domain in such a way that the fill distance of the data to its domain tends to 0 (Cressie, 2015, §5.8). We say an estimator is *consistent for the test risk under infill asymptotics* if – for any Q^{test}, χ, and h satisfying our assumptions above – the estimator converges in probability to $R_{Q^{\text{test}}}(h)$.

Definition 3.2 (Consistency of Test Risk Estimation Under Infill Asymptotics)**.** Fix a predictive method h and a test measure Q^{test}. Take Assumptions 2.1, 2.2, 2.4 and 2.5. Consider an infinite sequence of validation sets of increasing size: $(D_N^{\text{val}})_{N=1}^{\infty}$, $D_N^{\text{val}} = (S_n^{\text{val}}, X_n^{\text{val}}, Y_n^{\text{val}})_{n=1}^{N}$ such that when $N' < N$, the first N' points of D_N^{val} are $D_{N'}^{\text{val}}$. Suppose $\lim_{N\to\infty} \zeta(S_{1:N}^{\text{val}}, \mathcal{S}) = 0$. Let $\hat{R}_N$ be an estimator constructed from the validation data D_N^{val}. We say that the estimator $\hat{R}_N$ is consistent for the test risk under infill asymptotics if for all $\epsilon > 0, \lim_{N\to\infty} \Pr(|\hat{R}_N - R_{Q^{\text{test}}}(h)| \geq \epsilon) = 0$.

That is, as validation data fills the spatial domain, the estimator should converge to the test risk – no matter the composition of test sites. Our assumption that fill distance tends to zero is generally weaker than an assumption that the validation sites are drawn i.i.d. from a distribution with Lebesgue density supported on the spatial domain. Reznikov & Saff (2015, Theorem 2.1) showed an implication relationship between these assumptions in a much more general setting, and Vacher et al. (2021, Lemma 12) discuss the special case for the unit cube. Next, we present a finite-sample version of this implication, with an advantage relative to past work that we keep track of all constants.

Proposition 3.3 (Independent and Identically Distributed Data Satisfies an Infill Assumption)**.** *Suppose that* $\mathcal{S} = [0,1]^d$*,* $S_n^{\text{val}} \overset{\text{iid}}{\sim} P$ *for* $1 \leq n \leq N^{\text{val}}$*, and* P *has Lebesgue density lower bounded by* $c > 0$ *over* $[0,1]^d$*. Let* $B_d = \pi^{d/2}/\Gamma(d/2+1)$ *be the volume of the* d*-dimensional Euclidean unit ball. For any* $\delta \in (0,1)$ *there exists an* n_0 *such that for all* $N^{\text{val}} \geq n_0$ *with probability at least* $1 - \delta$

$$\zeta(S_{1:N^{\text{val}}}^{\text{val}}, [0,1]^d) \leq \left(\frac{4^d}{cN^{\text{val}}B_d}\left(\log \frac{6^d N^{\text{val}}}{B_d \delta}\right)\right)^{1/d}. \quad (2)$$

We prove Prop. 3.3 in App. D.2. The right side of this bound is $O((\log N^{\text{val}}/N^{\text{val}})^{1/d})$, and so the fill distance converges to zero in probability under these assumptions.

Consistency under infill asymptotics is a minimal desirable property. Like traditional consistency, we emphasize that Def. 3.2 is just a single check among many. For instance, often practitioners will be interested in extrapolation far from observed data, which is not modelled by infill asymptotics and will need to be considered separately. Our only supposition here is that we will generally prefer test-risk estimators that satisfy consistency under infill asymptotics to those that do not.

4. Current Estimators Exhibit Inconsistency

Even though consistency under infill asymptotics is a minimal desirable property, we next prove that principle existing test-risk estimators fail to satisfy it in realistic problems.

[3] The fill distance is not a distance in the mathematical sense since it is asymmetric and can equal 0 in cases when its two arguments are not exactly equal.

Inconsistency of the Holdout. We state our result and then discuss the realism of the example.

Proposition 4.1 (Inconsistency of holdout)**.** *There exists a set of test points and a data-generating process satisfying infill asymptotics such that $\hat{R}_{\text{Hold}}$ is not a consistent estimator of the test risk.*

While the holdout estimator of test risk is consistent for i.i.d. test and validation data (Devroye, 1976; Langford, 2005), we can construct examples showing Prop. 4.1 by observing that $\hat{R}_{\text{Hold}}$ has no dependence on the test task. So unless all test tasks have the same risk (which will be true only in unusually simplistic spatial settings), it cannot estimate them all consistently. The holdout estimator will generally exhibit non-trivial bias since it averages loss across the validation sites when we really care about loss at the test sites. See App. D.3.1 for a formal proof and also an example where the holdout converges to Δ, the maximum possible error under the loss bound.

Nearest Neighbor Estimator. Because of the assumed regularity in the error function (Assumption 2.5), it is natural to estimate the error at a test site using nearby validation points. Loog (2012) proposed risk estimators using k-nearest neighbors in the context of covariate shift. Both Loog (2012) and Portier et al. (2023) advocated for the use of 1-nearest neighbor (1NN) in the covariate-shift setting, with the latter providing theoretical justifications under standard covariate-shift assumptions. However, we show the 1NN estimator exhibits inconsistency in our spatial setting.

We first review a general k-nearest neighbor estimator, which we revisit later. Define the k-nearest neighbor radius of a point $S \in \mathcal{S}$ as $\tau^k(S) := \inf\{a \in \mathbb{R} : |S^{\text{val}}_{1:N^{\text{val}}} \cap B(S,a)| \geq k\}$, where $B(S,a)$ is the ball of radius a centered at S. The k-nearest neighbor set[4] of a point $S \in \mathcal{S}$ is $A^k(S) := \{1 \leq n \leq N^{\text{val}} : S^{\text{val}}_n \in B(S, \tau^k(S))\}$. As long as $1 \leq k \leq N^{\text{val}}$, $A^k(S)$ contains at least k points: the k nearest neighbors to S in the validation set. It may be larger than k if multiple points are equidistant from S.

Definition 4.2. The k-nearest neighbor (kNN) test-risk estimator is defined by $\hat{R}_{\text{NN},k}(h) := \sum_{n=1}^{N^{\text{val}}} w_n^{\text{NN},k} \ell(Y^{\text{val}}_n, h^{\chi}(S^{\text{val}}_n))$, where $w_n^{\text{NN},k} := (1/M^{\text{test}}) \sum_{m=1}^{M^{\text{test}}} 1\{S^{\text{val}}_n \in A^k(S^{\text{test}}_m)\}/|A^k(S^{\text{test}}_m)|$.

Loog (2012) proposed weighting the loss function in this way when training a model under covariate shift. Portier et al. (2023) analyzed a similar approach, in which validation points are sampled with probabilities corresponding to the weights in Def. 4.2, for mean estimation. Portier et al. (2023) made standard covariate shift assumptions of i.i.d. validation sites, i.i.d. test sites, and a bounded density ratio between the validation and test distributions.

[4]We will state our results for the version of nearest neighbors where ties are resolved by including all equidistant points. However, our analysis holds for arbitrary tie-breaking methods.

Inconsistency of 1NN. We again state our result and then develop intuition.

Proposition 4.3 (Inconsistency of 1NN)**.** *There exist a set of test points and a data-generating process satisfying infill asymptotics such that $\hat{R}_{\text{NN},1}(h)$ is not a consistent estimator of the test risk.*

For intuition, recall that—unlike in the covariate-shift setting—test points in the spatial setting are commonly fixed rather than arising i.i.d. from a distribution. Consider the simple case where $Q^{\text{test}} = \delta_S$ for some $S \in \mathcal{S}$. Using $k = 1$ leads us to estimate the error using a single validation point, which is inconsistent due to observation noise at the validation point. Where the problem with the holdout estimator was bias, the problem with 1NN is variance. In App. D.3.2 we prove Prop. 4.3 and show 1NN has large error when applied to classification point prediction tasks.

Inconsistency of Nearest Neighbors When k Is a Function of the Number of Validation Points. In fact, we can prove a more general result: that any nearest-neighbor test-risk estimator where the number of neighbors depends (only) on the number of validation points is inconsistent under infill asymptotics, regardless of type of dependence.

Proposition 4.4 (Inconsistency of kNN depending on number of validation points)**.** *Let $(k_n)_{n=1}^{\infty}$ be any sequence of natural numbers. Define the sequence of estimators $\hat{R}_{N^{\text{val}}}$ to be the nearest neighbor risk estimators using N^{val} validation points and $k_{N^{\text{val}}}$ neighbors. Then there exists a data-generating process satisfying infill asymptotics, a test set containing a single point, a predictive method h resulting in an error function satisfying the Lipschitz assumption, and an $\epsilon, \delta > 0$ such that with probability at least $1-\delta$, $\forall N^{\text{val}}$, $|\hat{R}_{N^{\text{val}}}(h) - R_{Q^{\text{test}}}(h)| \geq \epsilon$.*

See App. D.3.5 for a proof. There are two cases. (1) If the number of neighbors is bounded, the estimator suffers from non-vanishing variance as in the 1NN case. Or (2) the number is unbounded, so there exists a sequence of validation sites that accumulates slowly enough around each test site to lead to non-vanishing bias. Inconsistency of both 1NN and the holdout can be seen as corollaries of Prop. 4.4; for 1NN, choose: $\forall n, k_n = 1$. For the holdout, choose: $\forall n, k_n = n$.

5. A Consistent Estimator

We next provide a novel bound on the test risk estimation error of kNN. We propose using a kNN estimator with k chosen by optimizing our bound. We show that our proposed estimator is consistent for test risk under infill asymptotics. We here focus on error estimation; in App. C we provide

promising results for model selection and discuss open challenges.

5.1. Our Bound and Estimator

In light of the examples in Section 4, we propose to trade off the larger variance of small k and larger bias of large k by optimizing a bound depending on the validation set. Crucially, we adapt k using the actual locations of the test and validation sites, as Prop. 4.4 suggests such adaptivity is necessary. To that end, we first derive a bound on the test-risk estimation error as a function of k and the locations of test and validation sites. To state our bound, it will be useful to define the *k^{th}-order fill distance*[5] of a set Ψ_1 in a set Ψ_2 as the maximum distance from a point in Ψ_2 to its k^{th} nearest neighbor in Ψ_1:

$$\zeta^k(\Psi_1; \Psi_2) = \sup_{S_2 \in \Psi_2} \inf_{A \subset \Psi_1, |A|=k} \sup_{S_1 \in A} d_{\mathcal{S}}(S_1, S_2). \quad (3)$$

Theorem 5.1 (Bound on Estimation Error in Terms of Fill Distance)**.** *Consider a validation set D^{val} of size N^{val} and a test set D^{test} of size M^{test}. Take the k-nearest neighbors test-risk estimator from Def. 4.2. Choose $\delta \in (0,1)$ and k such that $1 \leq k \leq N^{\text{val}}$. Let $\rho_k := \zeta^k(S^{\text{val}}_{1:N^{\text{val}}}, S^{\text{test}}_{1:M^{\text{test}}})$ and $\beta_\delta := \Delta\sqrt{\frac{1}{2}\log\frac{2}{\delta}}$. Take Assumptions 2.1, 2.2, 2.4 and 2.5. Then, with probability at least $1-\delta$,*

$$|R_{Q^{\text{test}}}(h) - \hat{R}_{\text{NN},k}(h)| \leq L\rho_k + \beta_\delta \|w^{\text{NN},k}\|_2 \quad (4)$$

$$\leq L\rho_k + \beta_\delta \sqrt{\max_{1 \leq n \leq N^{\text{val}}} Q^{\text{test}}(B(S^{\text{val}}_n, \rho_k))/k}, \quad (5)$$

where $B(S,r)$ denotes the ball of radius r centered at S. See Assumptions 2.1 and 2.5 and Def. 2.3 for Δ, L, Q^{test} respectively.

We prove Thm. 5.1 in App. D.4. We use the right-hand side of Eqn. (4) algorithmically, and the right-hand side of the Eqn. (5) to gain intuition for cases under which the bound is small, as well as in proofs. The first term on the far-right-hand side in Eqn. (5) is a worst-case upper bound on the bias of our estimator; it is large if the average loss varies quickly in space or if validation data is not available near test data. Larger k may increase the first term. The second term comes from applying a tail bound; if most of the weight is put on a few validation points, the resulting estimator has high variance and this term is large. Sufficiently large k will decrease the second term. If we can find a k such that both (a) the distance from each test point to its k nearest neighbors is small and (b) no validation point has too much impact on our estimator, $\hat{R}_{\text{NN},k}(h)$ provides a good estimate for $R_{Q^{\text{test}}}(h)$.

[5]We assume in this definition that all spatial locations are distinct. If not, Ψ_1 should be treated as a multi-set.

Thm. 5.1 is closely related to Portier et al. (2023, Prop. 4). While there are technical differences in the proof and algorithm (and the risk that is bounded), the substantive distinction is that we state our bound directly in terms of the fill distance, instead of upper bounding this distance again as done in Portier et al. (2023). We can therefore avoid making assumptions about the distributions of the sites; we instead highlight the fill distance of the validation set as an essential quantity in controlling the accuracy of nearest neighbor risk estimation.

Selection Procedure with Unknown Lipschitz Constant. If the Lipschitz constant of the average loss, L, can be upper bounded, for example by knowledge about how quickly varying the spatial processes involved in the analysis are, then k can be selected by minimizing the first upper bound in Eqn. (5). Since the bound is conditional on the validation and test sites, the bound still holds with the same probability for k selected by this minimization. However, it will generally be the case that the Lipschitz constant is unknown. We therefore suggest choosing the number of neighbors by minimizing the upper bound from Thm. 5.1 with 1 in place of the Lipschitz constant:

$$k^\star_T \in \arg\min_{k \in T} \rho_k + \beta_\delta \|w^{\text{NN},k}\|_2. \quad (6)$$

For computational efficiency, we focus on choosing k as a power of 2: $T = T_2 := \{2^i\}_{i=1}^{\lfloor \log_2 N^{\text{val}} \rfloor}$. We call the resulting estimator *spatial nearest neighbors* (SNN).

5.2. Our Method is Consistent

We show that SNN is consistent under infill asymptotics.

Corollary 5.2 (Our Method is Consistent under Infill Asymptotics)**.** *Let $\mathcal{S} = [0,1]^d$. Take Assumptions 2.1, 2.2, 2.4 and 2.5. Let $\tilde{\rho} := \zeta(S^{\text{val}}_{1:N}, \mathcal{S})$. Let $k^\star_{T_2} \in \arg\min_{k \in T_2} \rho_k + \beta_\delta \|w^{\text{NN},k}\|_2$ with $\delta = \min(1, r)$ and $r \in [c\tilde{\rho}, C\tilde{\rho}]$ for some constants (possibly depending on dimension) $c, C > 0$. Then the $k^\star_{T_2}$-nearest neighbor risk estimator is consistent under infill asymptotics.*

See App. D.5 for a proof of Cor. 5.2. Cor. 5.2 states that selecting the number of neighbors by minimizing an upper bound on our error in estimation leads to an estimator that is consistent regardless of the test data, as long as the validation data are dense on the unit cube. In App. D.5.4, we provide a computationally efficient algorithm for calculating an r satisfying the condition $c\tilde{\rho} \leq r \leq C\tilde{\rho}$, and we prove the correctness of this algorithm.

6. Experiments

Our theory suggests the holdout exhibits substantial bias in many tasks. And we expect 1NN to exhibit substantial variance in many point prediction tasks. Our experiments

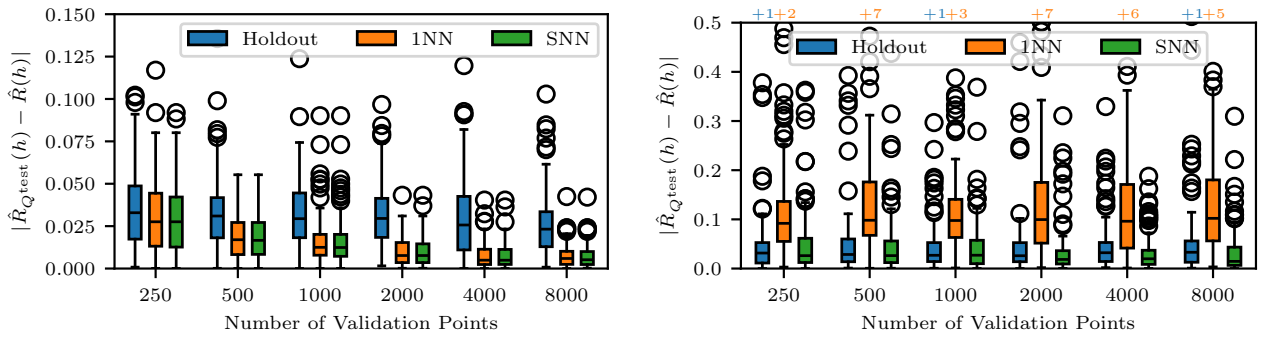

Figure 1: Error for test risk estimation in the *grid prediction task (left)* and *point prediction task (right)* across methods (holdout in blue, 1NN in orange, our SNN in green); lower values correspond to better performance. The vertical axis shows the absolute difference between the estimated test risk and empirical test risk. Each box plot shows the median, inter-quartile range, and outliers based on 100 synthetic datasets. The horizontal axis tracks increasing validation set sizes. Numbers above the upper box indicate the number of outliers falling above the vertical limit.

confirm these observations. While there exist tasks where either the holdout or 1NN performs similarly to SNN, there are also many tasks where each performs much worse than SNN. Since SNN performs well across all experiments, we prefer SNN when a new task arises.

Ground Truth. In a traditional machine learning prediction task, analysts ask how well a predictive method predicts the observed response at a set of covariates, so the observed response would form the ground truth. Since here we instead judge evaluation methods, we must ask how well the evaluation method estimates test risk (Def. 2.3); that is, the true test risk now forms ground truth. From Assumption 2.4, the true test risk requires an integral over the (unknown) noise P_ϵ. Accessing ground-truth responses is often easy; by contrast, it is highly unusual to access even a high-quality approximation of the integral for test risk (much less the exact test risk) in a real task. Therefore we devise a series of workarounds. First, we consider realistic tasks with simulated data. Second, we consider a realistic task with a semi-simulated data set, where we control the noise distribution by constructing it from bootstrapped residuals that arise from real data. Third, we use fully real data to construct a ground truth by considering an unrealistic task. Finally, we consider fully real data and a realistic task by forfeiting access to ground truth.

6.1. Test Risk Estimation on Fully Synthetic Data

We set up two fully synthetic experiments: a grid prediction task and a point prediction task. Based on our analyses above, we expect the holdout to struggle with the former and 1NN to struggle with the latter; our experiments confirm this intuition. See App. E.3 for full experiment details.

Validation Data, Test Data, and Ground Truth. In both experiments, we vary $N^{\text{val}} \in \{250, 500, 1000, 2000, 4000, 8000\}$. We use a truncated squared loss: $\ell(a, b) = \max((a-b)^2, 1.0)$. For the *grid task*, the test sites comprise a 50×50 grid of equally spaced points in $[-0.5, 0.5]^2$ (orange points in Fig. 3). We generate the validation sites via a sequential process that leads to clustering (blue points in Fig. 3). For the *point task*, there is a single test site at $(0, 0)$. Validation sites are i.i.d. uniform in $[-0.5, 0.5]^2$. For both tasks, we generate covariates and responses conditional on the sites: $Y_i^j = f(\chi^1(S_i^j), \chi^2(S_i^j)) + \eta(S_i^j) + \epsilon_i^j$, $\epsilon_i^j \overset{\text{iid}}{\sim} N(0, \sigma^2)$. We generate η, χ^1, χ^2, f according to independent Gaussian processes (GPs); we describe our kernel and parameter choices in App. E.3.1. We plot examples of the generated data in Figs. 4 and 6. We make draws from the data-generating process to form an unbiased Monte Carlo estimate of the test risk, $\hat{R}_{Q^{\text{test}}}(h)$, and use this estimate as ground truth; see App. E.3.5 for details.

Spatial Predictive Method. To arrive at our spatial predictive method, we generate training data according to the same distribution as the validation data. Since real-world analyses are often missing potentially relevant covariates, we retain only the first covariate (and not the second) as a realistic form of misspecification. We fit a GP regression, with zero prior mean and the same kernel used in data generation; we predict using the posterior mean.

Results. We expect our SNN estimator to be consistent in all tasks. In the grid task, we expect the variance of 1NN to be low since there are many test points spread across the domain. And we expect the bias of the holdout to be high since the test and validation points have noticeably different spatial arrangement. Our results in the grid task (Fig. 1, left) agree with our intuitions; the errors of 1NN and our SNN decrease much more rapidly across the values of N^{val} than

the holdout.

Given the single test point in the point task, we expect the high variance of 1NN to be an issue with substantial probability. Fig. 1 (right) agrees with our intuition; the error of our SNN estimator decreases much more rapidly across the values of N^{val} than 1NN. In this case, we find that the holdout errors decrease rapidly as well. We also plot the (signed) relative difference of each estimator to the empirical test risk in Figs. 5 and 7 (App. E.3).

In Tables 2 and 3 (App. E.3), we show k, the number of nearest neighbors selected by SNN for each of the two tasks. In the grid task, the k selected was at most 4 in all cases we considered. In the point task, the value of k selected increased with N^{val}, though it always remained over an order of magnitude smaller than N^{val}.

6.2. Air Temperature Task with Bootstrapped Residuals

We next consider a real task on a semi-synthetic dataset. We find that 1NN performs poorly; while the holdout performs best, SNN performs well. Full details can be found in App. E.4.

Data and Ground Truth. Our test task is prediction of monthly average air temperature in January 2023 at the 5 largest urban areas in the United States (New York City, Los Angeles, Chicago, Miami, and Houston), based on available weather station data in the same month (Menne et al., 2018). Loss is truncated absolute error (in °C). To access ground truth test risk, we create a partially synthetic response variable. We first fit Gaussian process regression (GPR) to all the available weather station data. We build 100 synthetic datasets by calculating the residuals of the posterior mean of this model, sampling a residual value for each weather station and point we want to predict at and adding these to the mean prediction of the model. Because we then have access to samples from the distribution of response values at the test sites under this data-generating process, we can obtain a (accurate estimate of) ground truth test risk of a predictive method on this partially synthetic response (App. E.4.7).

Spatial Predictive Methods. We train two predictive methods on this data: GPR and a geographically weighted regression (GWR) based on MODIS-Aqua (Wan et al., 2021) land surface temperature measurements, inspired by Hooker et al. (2018). We use 50% of the weather station locations for training the predictive methods and the remaining 50% for validation (3211 observations in each).

Results. The error in estimating the predictive performance of both methods is shown in Fig. 2 (far and mid left) for 100 different datasets with different samples of the residuals (but the same training and validation split). Given the point prediction task, we expect 1NN (orange, middle) to have a high variance; the figure confirms our intuition. The estimates given by SNN (green, right) and the holdout (blue, left) are much closer to the ground truth. In this case, the holdout has a small bias, and its variance is substantially lower than SNN. So, in this case, the holdout typically returns slightly better estimates of the error than SNN. If many more weather stations were used for validation, we expect that the SNN would eventually outperform the holdout estimate.

6.3. Property Sales in England and Wales

Here and in Section 6.4, we define somewhat unrealistic tasks to access ground truth on fully real data. Here we find that 1NN and SNN perform similarly well while the holdout exhibits a large bias. App. E.5 contains additional details and figures for this experiment.

Data and Ground Truth. We consider prediction of the price of a flat in England and Wales based on location, loosely following Hensman et al. (2013) but using data from 2023 (HM Land Registry, 2023). We make 100 datasets by sampling a training dataset of 40,000 points from flat sales outside London, a test set consisting of 1,000 flat sales within London, and a validation set consisting of the remaining sales in 2023 (31,484 outside London, 21,179 in London). The loss is truncated mean absolute error. In App. E.5.3, we justify how we can form a high-quality estimate of ground-truth test risk by: assuming a form of independence, using a bounded loss, and applying Hoeffding's inequality.

Spatial Predictive Method. We fit a Gaussian process regression model with variational inference as in Hensman et al. (2013) with minor modifications; we use a sum of two Matérn 3/2 kernels. We use the non-stochastic version of variational inference in GPR (Titsias, 2009) to avoid known difficulties with tuning hyperparameters in the stochastic version (Ober et al., 2024), and 2000 inducing points.

Results. Because the model is trained using data only outside London, we expect it to make larger errors predicting flat sales in London than outside London. As a result, we expect the holdout to have a large bias. We expect both 1NN and SNN to perform similarly: since the test sites are sampled randomly within London, we expect them to have different nearest neighbors and so variance of 1NN should be reasonably small. The mean absolute error of each method, relative to the estimate of the ground truth, is shown in Fig. 2 (mid right). As expected, the holdout substantially underestimates the test risk, while the other estimators perform reasonably well.

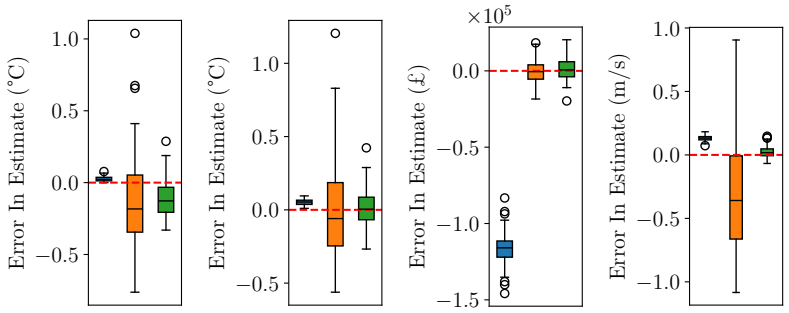

Figure 2: Signed errors in estimating the test risk for (left to right) the bootstrapped experiment with geographically weighted regression; the same task with GP regression; the flat price prediction task; and the wind speed prediction task. The holdout (blue), 1NN (orange), and SNN (green) appear left to right in each plot.

Table 1: Test risk estimates for the 5-metros task. Rows correspond to predictive methods and columns to estimators. We report two standard error intervals for holdout. For each estimator, we bold the predictive method with lower estimated test risk.

	GWR	Spatial GP
Hold.	**0.83 ±0.03**	**0.90 ±0.04**
1NN	0.61	**0.44**
SNN	**0.53**	0.61

6.4. Wind Speed Prediction

In this next experiment with an unrealistic task but fully real data, we find that both the holdout and 1NN perform poorly while SNN performs well. Full details are in App. E.6.

Data and ground truth. Our test task is predicting the average wind speed on a typical day in January at Chicago O'Hare airport, from daily historical weather station data (Menne et al., 2012). There are 775 historical weather station observations at the Chicago O'Hare site in January months. We split the remaining weather stations (at a station level) into a training and validation set consisting of 962 training stations and 241 validation stations. Each station has a different number of observations depending on how complete the historical record of average daily wind speed data is at this site. Typically the training set consists of on the order of 580,000 measurements and the validation set around 126,000 measurements. We perturb the location labels of each measurement at each weather station by a tiny amount to avoid ties when running nearest neighbors. The test set contains all of January rather than just a particular date so that we can form a high-quality estimate of ground truth. To form ground truth, we also assume that average wind speed decorrelates in time quickly. The loss is truncated (root) mean square error. See App. E.6.3 for more details.

Spatial Predictive Method. We use LightGBM (Ke et al., 2017) to make predictions.

Results. We expect 1NN to have a high variance because the task is point prediction. Fig. 2 (right) confirms this. The holdout has large bias for this task while SNN exhibits low bias and low variance.

6.5. Air Temperature Prediction with Real Response

We finally consider a case with real data and a real task. Although we cannot access ground truth, we show that the holdout, 1NN, and our SNN give very different estimates of test risk and can differ in model selection. Given all the previous results, we suggest using SNN. See App. E.4 for full details and also a grid-prediction task where all estimators are in agreement.

Data and Models Fit. The data, test task and models considered are the same as in Section 6.2, but the actual response values are used for training and validation.

Results. Table 1 shows a large discrepancy in test risk estimates across estimators. We see that 1NN chooses a different predictive method (spatial GP) than the holdout or our SNN does.

Acknowledgements

This work was supported in part by an NSF CAREER Award and the Office of Naval Research under grant N00014-20-1-2023 (MURI ML-SCOPE).

References

Banzon, V., Smith, T. M., Chin, T. M., Liu, C., and Hankins, W. A long-term record of blended satellite and in situ sea-surface temperature for climate monitoring, modeling and environmental studies. *Earth System Science Data*, 8 (1):165–176, 2016.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816 – 845, 2023.

Barbet-Massin, M., Rome, Q., Villemant, C., and Courchamp, F. Can species distribution models really predict

the expansion of invasive species? *PloS one*, 13(3): e0193085, 2018.

Bates, S., Hastie, T., and Tibshirani, R. Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association*, 2023.

Burman, P., Chow, E., and Nolan, D. A cross-validatory method for dependent data. *Biometrika*, 81(2):351–358, 1994.

Burt, D. R., Rasmussen, C. E., and van der Wilk, M. Convergence of sparse variational inference in gaussian processes regression. *Journal of Machine Learning Research*, 21(131):1–63, 2020. URL http://jmlr.org/papers/v21/19-1015.html.

Cressie, N. *Statistics for spatial data*. John Wiley & Sons, 2015.

De Bruin, S., Brus, D. J., Heuvelink, G. B., van Ebbenhorst Tengbergen, T., and Wadoux, A. M.-C. Dealing with clustered samples for assessing map accuracy by cross-validation. *Ecological Informatics*, 2022.

Devroye, L. *Nonparametric discrimination and density estimation*. PhD thesis, University of Texas at Austin., 1976.

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., Patton, B., Alemi, A., Hoffman, M., and Saurous, R. A. Tensorflow distributions. *arXiv preprint arXiv:1711.10604*, 2017.

Duan, S.-B., Li, Z.-L., Li, H., Göttsche, F.-M., Wu, H., Zhao, W., Leng, P., Zhang, X., and Coll, C. Validation of Collection 6 MODIS land surface temperature product using in situ measurements. *Remote sensing of environment*, 225:16–29, 2019.

Foster, L., Waagen, A., Aijaz, N., Hurley, M., Luis, A., Rinsky, J., Satyavolu, C., Way, M. J., Gazis, P., and Srivastava, A. Stable and efficient Gaussian process calculations. *Journal of Machine Learning Research*, 10(4), 2009.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 2009.

Gupta, P., Remer, L. A., Levy, R. C., and Shana, M. Validation of MODIS 3 km land aerosol optical depth from NASA's EOS Terra and Aqua missions. *Atmospheric Measurement Techniques*, 2018.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

Hensman, J., Fusi, N., and Lawrence, N. D. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence*, pp. 282, 2013.

HM Land Registry. Price paid data, 2023. URL https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads. Contains HM Land Registry data © Crown copyright and database right 2021. This data is licensed under the Open Government Licence v3.0.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 1963.

Hooker, J., Duveiller, G., and Cescatti, A. A global dataset of air temperature derived from satellite remote sensing and weather stations. *Scientific Data*, 5(1):180246, 2018.

(https://mathoverflow.net/users/36721/iosif pinelis), I. P. A problem on rate of decay of fill distance? MathOverflow, 2021. URL:https://mathoverflow.net/q/391340 (version: 2021-05-20).

Huangfu, Q. and Hall, J. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10, 03 2015.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Kianian, B., Liu, Y., and Chang, H. H. Imputing satellite-derived aerosol optical depth using a multi-resolution spatial model and random forest for PM2.5 prediction. *Remote Sensing*, 13(1), 2021.

Kpotufe, S. and Martinet, G. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 2021.

Langford, J. Tutorial on practical prediction theory for classification. *Journal of machine learning research*, 6 (3), 2005.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., and Bretagnolle, V. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography*, 23:811–820, 2014.

Lieske, D. and Bender, D. A robust test of spatial predictive models: Geographic cross-validation. *Journal of Environmental Informatics*, 17(2), 2011.

Linnenbrink, J., Milà, C., Ludwig, M., and Meyer, H. kNNDM: k-fold nearest neighbour distance matching cross-validation for map accuracy estimation. *EGUsphere*, 2023.

Loog, M. Nearest neighbor-based importance weighting. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6. IEEE, 2012.

Mahoney, M. J., Johnson, L. K., Silge, J., Frick, H., Kuhn, M., and Beier, C. M. Assessing the performance of spatial cross-validation approaches for models of spatially structured data. *arXiv preprint arXiv:2303.07334*, 2023.

Mao, H., Martin, R., and Reich, B. J. Valid model-free spatial prediction. *Journal of the American Statistical Association*, 2022.

Matthews, A. G. d. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z., and Hensman, J. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, 2017.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G. An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29(7):897 – 910, 2012.

Menne, M. J., Williams, C. N., Gleason, B. E., Rennie, J. J., and Lawrimore, J. H. The global historical climatology network monthly temperature dataset, version 4. *Journal of Climate*, 2018.

Meyer, H. and Pebesma, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 2021.

Meyer, H. and Pebesma, E. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, 2022.

Milà, C., Mateu, J., Pebesma, E., and Meyer, H. Nearest neighbour distance matching leave-one-out cross-validation for map validation. *Methods in Ecology and Evolution*, 2022.

Minnett, P. J. The validation of sea surface temperature retrievals from spaceborne infrared radiometers. In *Oceanography from Space: Revisited*. Springer Dordrecht, 2010.

Ober, S. W., Artemev, A., Wagenländer, M., Grobins, R., and van der Wilk, M. Recommendations for baselines and benchmarking approximate gaussian processes, 2024.

Office Of National Statistics. National statistics postcode lookup UK, 2024. Contains OS data © Crown copyright and database right 2024. This data is licensed under the Open Government Licence v3.0.

Okabe, A., Boots, B., Sugihara, K., and Chiu, S. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Wiley, 01 2000.

Omohundro, S. M. Five balltree construction algorithms. Technical report, International Computer Science Institute, UC Berkeley, 2009.

Özkaynak, H., Baxter, L. K., Dionisio, K. L., and Burke, J. Air pollution exposure prediction approaches used in air pollution epidemiology studies. *Journal of Exposure Science & Environmental Epidemiology*, 23(6):566–572, 2013.

Pathak, R., Ma, C., and Wainwright, M. J. A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, 2022.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

Pitman, J. *Combinatorial Stochastic Processes*. Springer, 2006.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C. F., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A. I., Gourlet-Fleury, S., and Pélissier, R. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 2020.

Portier, F., Truquet, L., and Yamane, I. Scalable and hyperparameter-free non-parametric covariate shift adaptation with conditional sampling. *arXiv preprint 2312.09969*, 2023.

Racine, J. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, 99(1):39–61, 2000.

Remigio, R. V., He, H., Raimann, J. G., Kotanko, P., Maddux, F. W., Sapkota, A. R., Liang, X.-Z., Puett, R., He, X., and Sapkota, A. Combined effects of air pollution and extreme heat events among ESKD patients within the northeastern united states. *Science of the Total Environment*, 2022.

Reznikov, A. and Saff, E. B. The covering radius of randomly distributed points on a manifold. *International Mathematics Research Notices*, 2016:6065–6094, 2015.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., and Thuiller, W. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 2017.

Sarafian, R., Kloog, I., Sarafian, E., Hough, I., and Rosenblatt, J. D. A domain adaptation approach for performance estimation of spatial predictions. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

Shabani, F., Kumar, L., and Ahmadi, M. A comparison of absolute performance of different correlative and mechanistic species distribution models in an independent area. *Ecology and Evolution*, 6(16):5973–5986, 2016.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 2000.

Telford, R. and Birks, H. The secret assumption of transfer functions: problems with spatial autocorrelation in evaluating model performance. *Quaternary Science Reviews*, 24(20):2173–2179, 2005.

Telford, R. and Birks, H. Evaluation of transfer functions in spatially structured environments. *Quaternary Science Reviews*, 28(13-14):1309–1316, 2009.

Thampi, A. reverse-geocoder Python package, 2015. URL https://github.com/thampiman/reverse-geocoder.

Tibshirani, R. J. Advanced topics in statistical learning: Minimax theory for nonparametric regression, 2023. URL https://www.stat.berkeley.edu/~ryantibs/statlearn-s23/lectures/minimax.pdf.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Titsias, M. Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pp. 567–574. PMLR, 2009.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

United States Census Bureau. Gazetteer files: Urban areas, 2023.

Vacher, A., Muzellec, B., Rudi, A., Bach, F., and Vialard, F.-X. A dimension-free computational upper-bound for smooth optimal transport estimation. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pp. 4143–4173, 2021.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2):225–232, 2019.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. Modelling species presence-only data with random forests. *Ecography*, 44(12):1731–1742, 2021.

Van Rossum, G. and Drake, F. L. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Wan, Z., Hook, S., and Hulley, G. MODIS/Aqua land surface temperature/emissivity monthly L3 global 0.05Deg CMG v061 (data set), 2021.

Wang, W., Fecht, D., Beevers, S., and Gulliver, J. Predicting daily concentrations of nitrogen dioxide, particulate matter and ozone at fine spatial scale in Great Britain. *Atmospheric Pollution Research*, 13(8):101506, 2022.

Wendland, H. *Scattered Data Approximation*. Cambridge University Press, 2004.

Werner, M., Kryza, M., and Guzikowski, J. Can data assimilation of surface PM2.5 and satellite AOD improve WRF-Chem forecasting? A case study for two scenarios of particulate air pollution episodes in Poland. *Remote Sensing*, 11(20), 2019.

Yu, Y.-L. and Szepesvári, C. Analysis of kernel mean matching under covariate shift. In *International Coference on International Conference on Machine Learning*, 2012.

Appendix

Table of Contents

A. Broader Impact Statement

This paper identifies potential failure modes of existing validation techniques in a spatial setting and suggests practical improvements, with supporting theory and empirics. We hope that our work can play a part in improving the reliability of prediction assessment and thereby help improve the credibility of scientific analyses.

B. Extended Related Work

B.1. Validation versus Cross-Validation

While we expect our work to have implications for cross validation (CV), we focus on validation here since (1) we do not assume training data exists or is easy to change (for example in the case of a large physical model) and (2) CV presents additional subtle challenges. The holdout is broadly applicable to any predictive method (whether data-driven, physical, or a combination thereof). However, in the cases when data is scarce and a data-driven predictive method is used, cross-validation is commonly believed to make more efficient use of the data. CV is widely used in spatial analyses; among many examples are Wang et al. (2022); Valavi et al. (2021); Kianian et al. (2021). However, the interpretation of the error estimates given by cross-validation is subtle even in the classical i.i.d. setting (Bates et al., 2023). We focus on the validation setting in this

work due to its broad applicability and clear interpretation, though extensions to cross-validation, as well as a theoretical understanding of how the resulting error estimate relate to predictive performance is a promising direction for future work.

B.2. Covariate Shift

In covariate shift, it is generally assumed that $S_n^{\text{val}} \overset{\text{iid}}{\sim} P^{\text{val}}$, $S_m^{\text{test}} \overset{\text{iid}}{\sim} Q$ and $U_i^{\ell}|S_i^{\ell} \overset{\text{iid}}{\sim} \mu$ for $\ell \in \{\text{val}, \text{test}\}$. Moreover, it is typically assumed the density ratio $\frac{dQ}{dP}$ exists and is bounded, although there is recent work relaxing this assumption in the context of non-parametric regression (Kpotufe & Martinet, 2021; Pathak et al., 2022). Mean estimation seeks to estimate $\mathbb{E}[U_1^{\text{test}}]$. Taking $(U_m^{\text{test}}|S_m^{\text{test}}) = \ell(Y_m^{\text{test}}, h^{\chi}(S_m^{\text{test}}))$, this is the same task we consider, but with the (stronger) assumption that covariates are independent and identically distributed with a bounded density ratio between the test and validation distributions. This assumptions is not appropriate for validation with many spatial datasets, with particularly simple examples being when the task of interest requires prediction at a single spatial location, or on a regular grid. Our assumptions are essentially a conditional formulation of mean estimation under covariate shift. Many methods proposed for addressing mean estimation under covariate shift, including all the approaches we describe below, are based on re-weighting validation points, then applying the holdout approach described earlier with these weights.

The kernel mean matching algorithm provides a solution under standard covariate shift assumptions with a bounded density ratio and assuming the average loss as a function of space e_h lives in a reproducing kernel Hilbert space (RKHS) and has small norm (Gretton et al., 2009). Yu & Szepesvári (2012) provide a finite sample bound for this method, showing that under the assumptions outlined above, with high probability kernel mean matching can estimate the $R_Q(h)$ with error $O_p(\frac{1}{\sqrt{M^{\text{test}}}} + \frac{1}{\sqrt{N^{\text{val}}}})$. Recently, Portier et al. (2023) considered the mean estimation problem under covariate shift, but relaxed the RKHS assumption to instead assume that the average loss is Lipschitz, the same assumption we take. Their estimator is built on nearest neighbor regression, and they advocate the use of 1-nearest neighbor. Compared to this work, the primary advantage of our analysis is that it removes the assumption that the sites are independent and identically distributed, making it directly relevant to tasks like grid prediction. Moreover, in the more general setting we consider, 1-nearest neighbor is not always consistent (Prop. 4.3), and using more neighbors can be beneficial. We provide a method for selecting the number of neighbors that has similar statistical properties as their approach for grid prediction, but retains consistency for a wider class of problems where using a single neighbor is no longer consistent.

B.2.1. Covariate Shift in the Context of Spatial Validation

Several recent works have applied the covariate shift framework to spatial problems. Sarafian et al. (2020) considered a weighted estimator motivated by importance weighting to address covariate shift. This suggests taking the weights to equal the density ratio of the test sites to the validation sites (which is assumed to exist) (Shimodaira, 2000). The density of the test sites was assumed to be known and a kernel density estimate was used to estimate the density of the validation sites. While Sarafian et al. (2020) observe this estimate is unbiased if the density ratio is known, in practice the estimator will be biased due to error in estimating the density ratio. While this estimator might be consistent with regularity assumptions on the densities, the assumption that the densities exist and have a bounded ratio is restrictive for many spatial tasks – for example, if we care about the quality of predictions at a few specific locations. De Bruin et al. (2022) consider a similar estimator to Sarafian et al. (2020), but normalize the weights to sum to 1 (which may not be the case for the weights given directly by density estimation).

B.3. Predictive Validation for Spatial Data

While our analysis focuses on variants of the holdout we also discuss methods for spatial cross-validation, as they can often be adapted to cases with held-out data.

B.3.1. Limitations of the Holdout Approach

The holdout has been empirically shown to under-estimate error for models trained with data more similar to the validation data than to the test data in several works; the review paper of Roberts et al. (2017) provides a detailed description of this phenomenon in the context of ecological statistics. Despite concerns raised in previous works, the holdout is widely used in spatial application areas for comparing methods and indicating the reliability of a given method.

B.3.2. Spatial Stratification Approaches

Concerns over the quality of the holdout estimate have led to the development of validation approaches based on evaluating the loss of a model on held-out data far from the data used to train the model (Telford & Birks, 2005). Cross-validation strategies for spatial datasets often also focus on evaluating a model on data that are far (in the spatial domain) from the data used to train the model. Spatially stratified blocking approaches are described in Lieske & Bender (2011) and Roberts et al. (2017), and several software packages make these cross-validation methods readily accessible in common statistical programming languages, especially `R` (Valavi et al., 2019; Mahoney et al., 2023). Similarly, variants of leave-one-out cross-validation in which points close to the point on which error will be assessed are also held-out during training have been developed for sequential data (Burman et al., 1994) and adapted to the spatial setting (Telford & Birks, 2009; Le Rest et al., 2014). Several simulation studies support claims that spatial buffering provides more realistic estimates of model risk than the standard cross-validation (Roberts et al., 2017; Mahoney et al., 2023). However, other simulation studies have shown that using spatially disjoint regions to train the model and to validate the model can lead to over-estimation of generalization error when the available data covers most of the space in which we are interested in making predictions (Ploton et al., 2020; De Bruin et al., 2022). While simulation studies show the strengths, and some of the limitations, of ensuring validation data are far in space from training data as a method for validating a predictive method, there is not clear theory establishing under what assumptions it allows for accurate evaluation of the risk, or consistent model selection. Roberts et al. (2017) offers some useful heuristics for when spatial stratification is preferable to the holdout. Racine (2000) provides a sketch for the consistency of the leave-one-out method described above for model selection in linear models with stationary sequential data, but we are not aware of a detailed proof clarifying the underlying assumptions about mixing of the process, extensions to non-linear models or extensions to the spatial setting.

B.3.3. Other Approaches for Spatial Validation

Other heuristics for estimating the error have emerged in the ecological statistics literature. De Bruin et al. (2022) also consider model based approaches based on non-parametric regression of the residual with a Gaussian process (kriging) to estimate the error of the model. The square of this regressor is then used as a plug-in estimator for estimating the squared loss. This approach is proposed as a heuristic, and it is not clear whether it is consistent, especially if the likelihood of the model fit to the residuals is misspecified, which will certainly be the case in practice. Milà et al. (2022); Linnenbrink et al. (2023) considered the distance between each validation site and its k-nearest neighbors in the training sites, and attempted to make the distribution of these distances similar to the distribution between the test sites and the training sites. However, it is not clear what assumptions on the data are needed for such a method to reliably estimate the generalization error of a method. Meyer & Pebesma (2021) emphasized that the validity of estimates of the generalization error of a spatial model depends on how similar the validation data are to the training data, relative to how similar test data are to the training data, and suggested only providing error estimates over an area that is judged to not be significantly more different from the training data than the available validation data are from the training data. The infill assumption we make in this work provides a particularly simple formalization of this idea, since it means that we have validation data close to every test point, and so we are able to reasonably estimate the error at each test point. Meyer & Pebesma (2022) provide a recent discussion of challenges of evaluating predictions made on a regular grid (map prediction) as well as other recent references for proposed spatial validations approaches. The point prediction problems we consider are analogous to the local error estimates they advocate for, while the grid prediction problems we consider are a form of global estimate.

B.3.4. Aspects of the Problem we do not Consider

If a statistical prediction method is used, there is a question of how to partition data between data used for training and validation, which is a central consideration in, for example Milà et al. (2022); Linnenbrink et al. (2023). In contrast, we focus on the case when a validation set is already decided upon. This allows our approach to be applied to physically-driven prediction methods, statistical methods and combinations thereof. Moreover, this allows our approach to be applied to models that have already been built when validation data becomes available and rebuilding the predictive method with a new training set would be expensive.

B.4. Non-exchangeable Conformal Prediction

Tools have been developed for providing confidence intervals for prediction with data that are not exchangeable using variants of conformal prediction (Tibshirani et al., 2019; Mao et al., 2022; Barber et al., 2023). Particularly relevant to our work is Mao et al. (2022), who construct confidence intervals at a specific location based on the error at its k-nearest

neighbors in the validation set. This is conceptually the same as the approach we take in mean estimation, but they focus on confidence intervals instead of risk estimation. They derive consistency results for the coverage of the intervals under an infill asymptotic setting, but do not show finite sample bounds which we prove in this work, making our results more quantitative. Finally, we give a theoretically-grounded method for choosing the number of neighbors by minimizing an upper bound on the error of the estimator, whereas it is unclear how to choose the number of neighbors in their approach.

C. Error Estimation for Model Selection

A practitioner will commonly select a predictive method within some collection by choosing the method with lowest estimated test risk. We consider two cases.

(1) Fixed predictive methods. So far we have focused on the setting where predictive methods are fixed in advance. As we accrue validation data, the consistency of SNN ensures it will eventually choose the predictive method with lowest test risk. Without consistency, holdout and 1NN cannot be trusted to choose the best predictive method.

(2) Proportional training and validation data. If we consider the special case where the predictive method is fit using training data (vs., say, a physical model), it is common for a practitioner to have a single set of available data that they then partition into training data and validation data; for instance, a fixed percentage of the total data may go to training. In this case, the predictive method changes as the validation set grows. Nonetheless, we can still be sure to eventually choose the predictive method with the lowest risk if the test-risk estimate has a faster rate of convergence in the number of validation data points than the convergence rate of the predictive method in the number of training points.

We are not able to formally characterize how SNN performs in model selection. But we provide rigorous results on rates of convergence of SNN that give suggestive guidance. In particular, (a) we first consider the case of a *grid prediction task*, where test sites are arranged on a grid. In this case, we are able to give a rate of convergence for both 1NN and SNN (Cors. D.16 and D.17). If the validation data are i.i.d. our rate is faster than the minimax optimal convergence rate for predicting a Lipschitz function in the presence of additive, homoskedastic noise. Our result shows the same rate of convergence as Portier et al. (2023, Prop. 4), who considered test data that was i.i.d. instead of on a grid. Our result is suggestive that both nearest neighbor methods may perform well at model selection for validating maps (grid prediction).

(b) Second, we provide finite-sample bounds and asymptotic characterizations for general (non-grid) Q^{test} (Cor. D.14). In this case, when validation data is i.i.d., up to logarithmic factors, we show that SNN converges *at* the optimal rate of convergence for Lipschitz functions (Cor. D.18). Since in this case, the SNN convergence rate is not *strictly* faster than the training convergence rate, it might be difficult to select between statistical methods that converge at the optimal regression rate. But we would still expect to be able to select between (i) a statistical method that converges at the optimal rate and (ii) one that does not (for example a misspecified parametric model). Since the holdout and 1NN may not even be consistent, we could not rely on them to select the better model even in this latter, easier case.

D. Proofs of Claims

We now present results and proofs not included in the main text. We essentially follow the order of results in the main text. Section App. D.1 gives sufficient conditions for Assumption 2.5 to hold for a homoskedastic, additive noise model and squared loss for responses taking values in $[0, 1]$, mentioned in Section 2.2. App. D.2 focuses on the proof of Prop. 3.3, first recalling several properties of covering numbers that will be used in the result. In App. D.3 we restate and prove our Props. 4.1, 4.3 and 4.4, which show limitations of existing validation methods in the spatial setting we study. In App. D.4 we prove a general result upper bounding the error in estimating the risk using k-nearest neighbors, as well as an upper bound on the error for $k^{\star}_{T_2}$. App. D.5 establishes the consistency of spatial nearest neighbors. App. D.6 discusses issues related to model selection and proves rates of convergence for spatial nearest neighbors.

D.1. Lipschitz Constant for Lipschitz Response and Predictive Method

While assuming the average loss is Lipschitz continuous as a function of space is mathematically convenient, it is perhaps more natural to make assumptions about the spatial field we are trying to make predictions about, as well as the predictive method we are using to make prediction. The following proposition gives an example of how Lipschitz continuity of the processes involved can imply Assumption 2.5.

Proposition D.1. *Consider $\mathcal{Y} \subset [0, 1]$ and squared loss. Suppose $f^{\chi}(S) := f(S, \chi(S))$ is L_Y-Lipschitz and h^{χ} is L_h-*

Lipschitz. Let (S, X, Y) be generated as in Assumption 2.2, with $Y = f(S, X) + \epsilon$. Then $e_h(S) \coloneqq \mathbb{E}[(Y - h^{\mathcal{X}}(S))^2|S]$ is $2(L_Y + L_h)$-Lipschitz.

Proof. Let $S, S' \in \mathcal{S}$ and ϵ, ϵ' be the associated noise random variables. Then,

$$|e_h(S) - e_h(S')| = \mathbb{E}[(f^{\mathcal{X}}(S) + \epsilon) - h^{\mathcal{X}}(S))^2 - (f^{\mathcal{X}}(S') + \epsilon') - h^{\mathcal{X}}(S'))^2] \tag{7}$$

$$= f^{\mathcal{X}}(S)^2 - f^{\mathcal{X}}(S)^2 + h^{\mathcal{X}}(S)^2 - h^{\mathcal{X}}(S')^2, \tag{8}$$

where we have used that $\mathbb{E}[\epsilon] = \mathbb{E}[\epsilon'] = 0$, ϵ, ϵ' are independent from h and $\mathbb{E}[\epsilon^2] = \mathbb{E}[(\epsilon')^2]$. Then,

$$f^{\mathcal{X}}(S)^2 - f^{\mathcal{X}}(S')^2 + h^{\mathcal{X}}(S)^2 - h^{\mathcal{X}}(S')^2 = (f^{\mathcal{X}}(S) + f^{\mathcal{X}}(S'))(f^{\mathcal{X}}(S) - f^{\mathcal{X}}(S')) \tag{9}$$

$$+ (h^{\mathcal{X}}(S) + h^{\mathcal{X}}(S'))(h^{\mathcal{X}}(S) - h^{\mathcal{X}}(S'))$$

$$\leq 2|f^{\mathcal{X}}(S) - f^{\mathcal{X}}(S')| + 2|h^{\mathcal{X}}(S) - h^{\mathcal{X}}(S')| \tag{10}$$

$$\leq 2(L_Y + L_h) d_{\mathcal{S}}(S, S'). \tag{11}$$

The first inequality uses that $\mathcal{Y} \subset [0, 1]$ and the second the Lipschitz assumptions on f and h. □

Therefore, at least in the case of squared loss with bounded response and predictive method values and a homoskedastic additive noise model, smoothness of the average response surface (as a function of space) together with smoothness of the predictive method imply Assumption 2.5. We expect to hold for other losses that are Lipschitz functions of the response and prediction.

D.2. Proof that Independent and Identically Distributed Data Implies Infill Asymptotic with High Probability

The purpose of this section is to prove Prop. 3.3. We begin by recalling this proposition:

Proposition 3.3 (Independent and Identically Distributed Data Satisfies an Infill Assumption)**.** *Suppose that $\mathcal{S} = [0, 1]^d$, $S_n^{\text{val}} \overset{\text{iid}}{\sim} P$ for $1 \leq n \leq N^{\text{val}}$, and P has Lebesgue density lower bounded by $c > 0$ over $[0, 1]^d$. Let $B_d = \pi^{d/2}/\Gamma(d/2 + 1)$ be the volume of the d-dimensional Euclidean unit ball. For any $\delta \in (0, 1)$ there exists an n_0 such that for all $N^{\text{val}} \geq n_0$ with probability at least $1 - \delta$*

$$\zeta(S_{1:N^{\text{val}}}^{\text{val}}, [0, 1]^d) \leq \left(\frac{4^d}{cN^{\text{val}}B_d}\left(\log \frac{6^d N^{\text{val}}}{B_d \delta}\right)\right)^{1/d}. \tag{2}$$

Our proof is similar to earlier proofs in Reznikov & Saff (2015, Section 5.2) and essentially follows the stackoverflow response (https://mathoverflow.net/users/36721/iosif pinelis) keeping track of numerical constants. Essentially, the idea is that were the fill distance to be large, there must be a ball of large radius that doesn't contain any points in the sample of location. But because the probability of a point in the sample falling in any ball is bounded below in terms of the volume of the ball, this must be improbable as the sample size grows.

D.2.1. Preliminary Definitions and Lemmas Related to Nets and Covering Number

In order to formalize the proof of Prop. 3.3 sketched above idea, we recall the definition of a net and covering number, as well as some standard properties of covering numbers.

Definition D.2 (Net, Covering Number)**.** Let $A \subset \mathbb{R}^d$ be a compact set. Any finite set $C \subset \mathbb{R}^d$ such that

$$A \subset \bigcup_{S \in C} B(S, \epsilon) \tag{12}$$

where $B(S, \epsilon)$ denotes the d-dimensional closed Euclidean ball of radius $\epsilon \geq 0$ centered at S is referred to as an ϵ-net of A. The ϵ-covering number of A, denoted by $\mathfrak{N}(\epsilon, A)$ is the minimum cardinality of an ϵ-net of A.

Because A is compact for any ϵ the ϵ-covering number of A is finite, and will generally increase as $\epsilon \to 0$.

We now recall how covering number behaves under affine transformations of the underlying space. For sets A, B and a scalar α, we define $A + B = \{a + b : a \in A, b \in B\}$ and $\alpha A = \{\alpha a : a \in A\}$.

Proposition D.3. *For any compact $A \subset \mathbb{R}^d$ and $c \in \mathbb{R}^d$, $\mathfrak{N}(\epsilon, A) = \mathfrak{N}(\epsilon, A + \{c\})$. For any $\alpha > 0$, $\mathfrak{N}(\epsilon, A) = \mathfrak{N}(\alpha\epsilon, \alpha A)$.*

Proof. The first claim is shown by noting that if C is an ϵ-net of A then $C+\{c\}$ is an ϵ-net of $A+\{c\}$ so that $\mathfrak{N}(\epsilon, A+\{c\}) \leq \mathfrak{N}(\epsilon, A)$. Applying a symmetric argument implies the reverse inequality.

For the second claim, let C be an ϵ-net for A. Then,

$$\alpha A \subset \alpha \left(\bigcup_{S \in C} B(S, \epsilon) \right) = \bigcup_{S \in C} B(\alpha S, \alpha\epsilon) = \bigcup_{S' \in \alpha C} B(S', \alpha\epsilon), \tag{13}$$

and so αC is an $\alpha\epsilon$-net of αA. It follows that $\mathfrak{N}(\alpha\epsilon, \alpha A) \leq \mathfrak{N}(\epsilon, A)$. The opposite inequality is obtained via the same argument applied with $\alpha' = \frac{1}{\alpha}$, $A' = \alpha A$ and $\epsilon' = \alpha\epsilon$. □

In the proof of Prop. 3.3 we apply a union bound over all elements in a net for the unit cube. We therefore need a result telling us that this net does not contain too many elements.

Lemma D.4 (Covering number of Unit Cube)**.** *Let* $\epsilon \in (0, 1]$. *Then* $\mathfrak{N}(\epsilon, [-1, 1]^d) \leq \frac{1}{B_d} \left(\frac{6}{\epsilon}\right)^d$

where $B_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ is the volume of the d-dimensional unit sphere.

Proof. By the upper bound in Wainwright (2019, Lemma 5.7),

$$\mathfrak{N}(\epsilon, [-1, 1]^d) \leq \frac{1}{B_d} \text{vol} \left(\left[-\frac{2}{\epsilon}, \frac{2}{\epsilon} \right]^d + B(0, 1) \right). \tag{14}$$

For any $S = S_1 + S_2 \in [-\frac{2}{\epsilon}, \frac{2}{\epsilon}]^d + B(0, 1)$ with $S_1 \in [-\frac{2}{\epsilon}, \frac{2}{\epsilon}]^d$ and $S_2 \in B(0, 1)$, by the triangle inequality for infinity norm,

$$\|S\|_\infty \leq \frac{2}{\epsilon} + 1 \leq \frac{3}{\epsilon}, \tag{15}$$

where we first used that points in the unit ball have infinity norm not more than 1 and then used that $\epsilon \leq 1$. Therefore,

$$\text{vol} \left(\left[-\frac{2}{\epsilon}, \frac{2}{\epsilon} \right]^d + B(0, 1) \right) \leq \text{vol} \left(\left[-\frac{3}{\epsilon}, \frac{3}{\epsilon} \right]^d \right) = \left(\frac{6}{\epsilon} \right)^d. \tag{16}$$

□

D.2.2. Main Proof

We again recall Prop. 3.3 for convenience when reading the proof.

Proposition 3.3 (Independent and Identically Distributed Data Satisfies an Infill Assumption)**.** *Suppose that* $\mathcal{S} = [0, 1]^d$, $S_n^{\text{val}} \overset{\text{iid}}{\sim} P$ *for* $1 \leq n \leq N^{\text{val}}$, *and* P *has Lebesgue density lower bounded by* $c > 0$ *over* $[0, 1]^d$. *Let* $B_d = \pi^{d/2}/\Gamma(d/2 + 1)$ *be the volume of the* d*-dimensional Euclidean unit ball. For any* $\delta \in (0, 1)$ *there exists an* n_0 *such that for all* $N^{\text{val}} \geq n_0$ *with probability at least* $1 - \delta$

$$\zeta(S_{1:N^{\text{val}}}^{\text{val}}, [0, 1]^d) \leq \left(\frac{4^d}{cN^{\text{val}} B_d} \left(\log \frac{6^d N^{\text{val}}}{B_d \delta} \right) \right)^{1/d}. \tag{2}$$

Proof of Prop. 3.3. For some $\tau \in (0, 1)$ (to be selected later) let C be a minimal cardinality $\tau/2$-net for $[0, 1]^d$. If $B(S, \tau/2)$ contains a validation point for all $S \in C$, then for any $S' \in [0, 1]^d$, by the triangle inequality,

$$\min_{1 \leq n \leq N^{\text{val}}} d_{\mathcal{S}}(S', S_n^{\text{val}}) \leq \min_{1 \leq n \leq N^{\text{val}}} \left(\min_{S' \in C} d_{\mathcal{S}}(S', S) + d_{\mathcal{S}}(S_n^{\text{val}}, S) \right) \leq \tau. \tag{17}$$

Therefore the probability that the fill distance is large ($> \tau$) is less than the probability that there exists an element of the net such that no validation point is close to it (within radius $\tau/2$):

$$\Pr(\zeta(S_{1:N^{\text{val}}}^{\text{val}}, [0, 1]^d) > \tau) \leq \Pr(\exists S \in C : S_n^{\text{val}} \notin B(S, \tau/2) \, \forall 1 \leq n \leq N^{\text{val}}). \tag{18}$$

The probability of any particular validation point falling in a ball centered at any point of radius contained in the $\tau/2$ can't be too small since P has density that is bounded below: For any $S \in [0,1]^d$

$$P(B(S,\tau/2)) = \int_{S' \in B(S,\tau/2)} \frac{dP}{d\lambda}(S') d\lambda(S') \geq \int_{S' \in B(S,\tau/2) \cap [0,1]^d} c d\lambda(S') \geq c \frac{\text{vol}(B(S,\tau/2))}{2^d}, \tag{19}$$

where λ denotes Lebesgue measure and in the final inequality we have used that since $S \in [0,1]^d$, at least one quadrant of $B(S,\tau/2)$ is contained in $[0,1]^d$.

Returning to Eqn. (18) and taking a union over all elements in the net:

$$\Pr(\zeta(S^{\text{val}}_{1:N^{\text{val}}}, [0,1]^d) > \tau) \leq \mathfrak{N}(\tau/2, [0,1]^d) \max_{a \in A} \Pr(S^{\text{val}}_n \notin B(a,\tau/2)\, \forall 1 \leq n \leq N^{\text{val}}) \tag{20}$$

$$\leq \mathfrak{N}(\tau/2, [0,1]^d) \max_{a \in A} (1 - P(B(a,\tau/2)))^{N^{\text{val}}} \tag{21}$$

$$\leq \mathfrak{N}(\tau/2, [0,1]^d) \left(1 - \frac{c}{2^d} B_d \left(\frac{\tau}{2}\right)^d\right)^{N^{\text{val}}}, \tag{22}$$

where $B_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ is the volume of the d-dimensional unit sphere. Eqn. (21) uses Eqn. (19). We now upper bound the terms in Eqn. (22).

Applying, Prop. D.3 and Lemma D.4

$$\mathfrak{N}(\tau/2, [0,1]^d) \leq \frac{1}{B_d} \left(\frac{6}{\tau}\right)^d. \tag{23}$$

Using the inequality $(1-x) \leq e^{-x}$,

$$\left(1 - \frac{c}{2^d} B_d \left(\frac{\tau}{2}\right)^d\right)^{N^{\text{val}}} \leq \exp\left(-\frac{c}{2^d} N^{\text{val}} B_d \left(\frac{\tau}{2}\right)^d\right). \tag{24}$$

Combining Eqns. (22) to (24),

$$\Pr(\zeta(S^{\text{val}}_{1:N^{\text{val}}}, [0,1]^d) > \tau) \leq \frac{1}{B_d} \left(\frac{6}{\tau}\right)^d \exp\left(-\frac{c}{2^d} N^{\text{val}} B_d \left(\frac{\tau}{2}\right)^d\right) \tag{25}$$

Now choose $\tau^d = \frac{4^d}{cN^{\text{val}} B_d} \log \frac{6^d N^{\text{val}}}{B_d \delta}$. For all N^{val} larger than some n_0, $\tau < 1$ because $\lim_{N^{\text{val}} \to \infty} \frac{4^d}{cN^{\text{val}} B_d} \log \frac{6^d N^{\text{val}}}{B_d \delta} = 0$, and so this choice satisfies our earlier assumption. For this choice of τ

$$\frac{1}{B_d} \left(\frac{6}{\tau}\right)^d \exp\left(-\frac{c}{2^d} N^{\text{val}} B_d \left(\frac{\tau}{2}\right)^d\right) = \frac{1}{\tau^d N^{\text{val}}} \delta = \frac{c B_d \delta}{4^d \log \frac{6^d N^{\text{val}}}{B_d \delta}} \leq \delta \frac{c}{2^d \log \frac{3^d N^{\text{val}}}{\delta}}, \tag{26}$$

where in the final inequality we use that $B_d \leq 2^d$ since the unit ball is contained in the unit cube. For all $N^{\text{val}} \geq \frac{\delta e^{c/2^d}}{3^d}$ the right-hand side is less than δ; but $\delta < 1, c \leq 1$ and so this holds for all $N^{\text{val}} \geq 1$. □

D.3. Proof of Inconsistency of Existing Methods

In this section, we restate and prove our results on limitations of existing methods for validation. These results were stated in Section 4 in the main text.

D.3.1. Inconsistency of the Holdout

We begin by focusing on the holdout. As sketched in the main text, the holdout does not depend on the particular test set, and therefore cannot approximate the test risk well for point prediction tasks unless the average loss at both points is the same.

Proposition 4.1 (Inconsistency of holdout)**.** *There exists a set of test points and a data-generating process satisfying infill asymptotics such that $\hat{R}_{\text{Hold}}$ is not a consistent estimator of the test risk.*

We prove the strong result.

Proposition D.5 (Counterexample to the consistency of holdout)**.** *Consider a single test point (S, X, Y) satisfying Assumption 2.2. Suppose the test risk at such a point (Def. 2.3 with $M^{\text{test}} = 1$) is not constant as a function of S. Then there exists a test set (of size one) such that $\hat{R}_{\text{Hold}}$ is not a consistent estimator of the test risk on that test set.*

Proof. Because $e_h(S)$ is not a constant $\mathcal{S}$ contains at least two elements S, S' such that $e_h(S) \neq e_h(S')$. Let $Q = \delta_S$ and $Q' = \delta_{S'}$. There there exists some $\gamma > 0$ such that

$$|R_Q(h) - R_{Q'}(h)| = |e_h(S) - e_h(S')| > 2\gamma. \tag{27}$$

If $\hat{R}_{\text{Hold}}$ is not consistent for $R_Q(h)$, we are done. Otherwise, by the definition of consistency, for all $N \geq N_0$ with probability at least $1/2$,

$$|(\hat{R}_{\text{Hold}})^{(N)}(h) - R_Q(h)| < \gamma. \tag{28}$$

where we use $(\hat{R}_{\text{Hold}})^{(N)}$ to denote the estimator constructed using the first N validation points, D_N^{val}. By the reverse triangle inequality,

$$|(\hat{R}_{\text{Hold}})^{(N)}(h) - R_{Q'}(h)| \geq |R_Q(h) - R_{Q'}(h)| - |(\hat{R}_{\text{Hold}})^{(N)}(h) - R_{Q'}(h)|. \tag{29}$$

Combining Eqn. (28) and Eqn. (29), for all $N \geq N_0$ with probability at least $1/2$

$$|(\hat{R}_{\text{Hold}})^{(N)}(h) - R_{Q'}(h)| > 2\gamma - \gamma = \gamma, \tag{30}$$

which implies the holdout is not consistent for $R_{Q'}(h)$. □

Proposition D.6. *Let $\ell(a, b) = \max(1, |a - b|)$. There exists a data-generating process, test set containing a single site and predictive method satisfying infill asymptotics such that the holdout converges in probability to 0, while $R_{Q^{\text{test}}} = 1$.*

Proof. Consider $\mathcal{S} = [0, 1]$, $h \equiv 0$, $Y = S$, $S^{\text{test}} = 1$ and

$$S_m^{\text{val}} = \begin{cases} U_m & m \text{ is prime,} \\ 0 & m \text{ otherwise.} \end{cases} \tag{31}$$

with U_m independent and identically distributed uniform variables. By the infinitude of primes, there are infinitely many m such that S_m^{val} is uniformly distributed and, for example by Prop. 3.3, this implies that S_m^{val} satisfies the infill assumption. On the other hand, by the prime number theorem, as $M^{\text{test}} \to \infty$, the density of primes in the natural numbers tends to 0, and so

$$\hat{R}_{\text{Hold}}(h) = \frac{1}{M^{\text{test}}} \sum_{m=1}^{M^{\text{test}}} S_m^{\text{val}} = \frac{1}{M^{\text{test}}} \sum_{\substack{m=1 \\ \text{prime}}}^{M^{\text{test}}} U_m \leq \frac{|\{m : 1 \leq m \leq M^{\text{test}}, m \text{ prime}\}|}{M^{\text{test}}} \to 0, \tag{32}$$

and $R_{Q^{\text{test}}}(h) = 1$. □

D.3.2. Inconsistency of 1-nearest neighbor Estimator

We now turn to 1-nearest neighbor risk estimation and restate and prove Prop. 4.3:

Proposition 4.3 (Inconsistency of 1NN)**.** *There exist a set of test points and a data-generating process satisfying infill asymptotics such that $\hat{R}_{\text{NN},1}(h)$ is not a consistent estimator of the test risk.*

We again actually prove a strong result

Proposition D.7. *Assume any test point satisfies Assumption 2.2. Assume there exists a constant $c > 0$ such that for any test point (S, X, Y), $\text{Var}[\ell(Y, h^X(S))|S] \geq c$. Next, consider an infinite sequence of validation sets as in Def. 3.2. Suppose there exists an $S' \in \mathcal{S}$ such that for any $r > 0$ and $N^{\text{val}} > 0$, $|\{1 \leq j \leq N^{\text{val}} : d_{\mathcal{S}}(S_j^{\text{val}}, S') = r\}| \leq 1$. Choose any Q^{test} such that $Q^{\text{test}}(\{S'\}) > 0$. Then there exists a $\delta \in (3/4, 1)$ and a $C^{(\delta)} > 0$ such that, for each N^{val}, with probability at least $1 - \delta$, $|R_{Q^{\text{test}}}(h) - (\hat{R}_{\text{NN},1})^{(N^{\text{val}})}(h)| \geq C^{(\delta)}$. Here $(\hat{R}_{\text{NN},1})^{(N^{\text{val}})}$ denotes the 1NN estimator associated to the first N^{val} data points and δ. $C^{(\delta)}$ and δ do not depend on N^{val} or other properties of the sequence of validation data.*

The technical condition $|\{1 \leq n \leq N^{\text{val}} : d_{\mathcal{S}}(S_j^{\text{val}}, S') = r\}| \leq 1$ ensures that the 1NN set for each point contains exactly 1 point. That is, there are no ties. This condition would be satisfied for the infinite sequence of validation sets with probability one if, for instance, the validation points were chosen i.i.d. from a uniform measure on a compact set; cf. Prop. 3.3. Alternatively, it can be removed if any form of tie-breaking that selects a single nearest neighbor is used in defining the estimator.

The idea of the proof is that if the loss has a non-vanishing variance, then the one nearest neighbor procedure results in an estimator with a non-zero variance for point prediction. Therefore, it cannot converge in probability to the actual risk, which is deterministic.

D.3.3. Preliminary Result

We begin by proving a result that says that if a bounded random variable has a second moment bounded below by C, then it cannot be close to zero most of the time. We will apply this inequality to the second moment of the difference between the 1-nearest neighbor estimator and the test risk to in our proof of Prop. 4.3.

Proposition D.8. *Let U be a random variable with $U \in [-A, A]$ almost surely and $\mathbb{E}[U^2] \geq C > 0$. Then for any $\delta \in (1 - \frac{C}{A^2}, 1)$ with probability $1 - \delta$*

$$|U| \geq A\sqrt{1 - \frac{1 - \frac{C}{A^2}}{\delta}}. \tag{33}$$

Proof. Because $U \in [-A, A]$, $A^2 - U^2$ is a non-negative random variable. Applying Markov's inequality, for any $t > 0$,

$$\Pr(A^2 - U^2 \geq t) \leq \frac{A^2 - \mathbb{E}[U^2]}{t} \leq \frac{A^2 - C}{t}. \tag{34}$$

Take $t = \frac{A^2 - C}{\delta}$ which is greater than 0 because $\delta \in (1 - \frac{C}{A^2}, 1)$. Then Eqn. (34) becomes,

$$\Pr\left(A^2 - U^2 \geq \frac{A^2 - C}{\delta}\right) \leq \delta. \tag{35}$$

Taking complements, with probability at least $1 - \delta$

$$A^2 - U^2 < \frac{A^2 - C}{\delta}. \tag{36}$$

Rearranging implies that with probability at least $1 - \delta$

$$U^2 > A^2 - \frac{A^2 - C}{\delta}. \tag{37}$$

For $\delta \in (1 - \frac{C}{A^2}, 1)$ this bound is non-vacuous (strictly greater than zero). Taking square roots, which is monotone, for any such δ with probability at least $1 - \delta$

$$|U| > A\sqrt{1 - \frac{1 - \frac{C}{A^2}}{\delta}}. \tag{38}$$

□

D.3.4. Proof of Inconsistency of 1-nearest neighbor

We return to our proof of Prop. 4.3. The idea will be to consider a point prediction task and then apply Prop. D.8 to show that with some fixed probability, the 1-nearest neighbor estimator is a fixed distance away from the test risk, even as N^{val} increases.

Proof of Prop. 4.3. Because expectation minimizes the squared error to a random variable over all constant functions

$$\mathbb{E}|R_{Q^{\text{test}}}(h) - (\hat{R}_{\text{NN},1})^{(N^{\text{val}})}(h)|^2 \geq \mathbb{E}|(\hat{R}_{\text{NN},1})^{(N^{\text{val}})}(h) - \mathbb{E}[(\hat{R}_{\text{NN},1})^{(N^{\text{val}})}(h)]|^2. \tag{39}$$

Because the ϵ_n^{val} are independent the variance of $(\hat{R}_{\text{NN},1})^{(N^{\text{val}})}$ is additive

$$\mathbb{E}|(\hat{R}_{\text{NN},1})^{(N^{\text{val}})}(h) - \mathbb{E}(\hat{R}_{\text{NN},1})^{(N^{\text{val}})}(h)|^2 = \sum_{n=1}^{N^{\text{val}}} (w_n^{\text{NN},1})^2 \mathbb{E}[\ell(Y_n^{\text{val}}, h^{\mathcal{X}}(S_n^{\text{val}})) - e_h(S_n^{\text{val}})] \tag{40}$$

$$\geq V \sum_{n=1}^{N^{\text{val}}} (w_n^{\text{NN},1})^2. \tag{41}$$

where $0 < V < \Delta^2/4$ is the assumed lower bound on the variance of $\ell(Y_n^{\text{val}}, h^{\mathcal{X}}(S_n^{\text{val}})) - e_h(S_n^{\text{val}})$ and we have left implicit the dependence of the weights on N^{val}. Also, $|\{1 \leq n \leq N^{\text{val}} : d_{\mathcal{S}}(S', S_n^{\text{val}}) = r| \leq 1$, implies S' has exactly one 1-nearest neighbor in $S_{1:N^{\text{val}}}^{\text{val}}$, call the index of this neighbor $n(S')$. Then

$$\sum_{n=1}^{N^{\text{val}}} (w_n^{\text{NN},1})^2 \geq (w_{n(S')}^{\text{NN},1})^2 \geq Q^{\text{test}}(\{S'\})^2. \tag{42}$$

Combining Eqn. (41) and Eqn. (42)

$$\mathbb{E}[|R_{Q^{\text{test}}}(h) - (\hat{R}_{\text{NN},1})^{(N^{\text{val}})}(h)|^2] \geq V Q^{\text{test}}(\{S'\})^2. \tag{43}$$

We now apply Prop. D.8 with $U = |R_{Q^{\text{test}}}(h) - (\hat{R}_{\text{NN},1})^{(N^{\text{val}})}(h)|$ to conclude that for $\delta \in (1 - \frac{VQ^{\text{test}}(\{S'\})^2}{\Delta^2}, 1)$ with probability at least $1 - \delta$

$$|R_{Q^{\text{test}}}(h) - (\hat{R}_{\text{NN},1})^{(N^{\text{val}})}(h)| \geq \Delta \sqrt{1 - \frac{1 - \frac{VQ^{\text{test}}(\{S'\})^2}{\Delta^2}}{\delta}} > 0. \tag{44}$$

As neither δ nor the right hand side of Eqn. (44) depend on N^{val}, 1-nearest neighbor is not consistent under infill asymptotics. □

Proposition D.9. *Consider $\mathcal{Y} = \{0, 1\}$ and $\ell(a, b) = \begin{cases} 0 & a = b \\ 1 & \text{otherwise} \end{cases}$, $\mathcal{S} = [0, 1]^d$ and $S_m^{\text{val}} \overset{\text{iid}}{\sim} \mu$, where μ is any measure with density with respect to Lebesgue measure. Fix any $S \in \mathcal{S}$ and any predictive method h and let $Q^{\text{test}} = \delta_S$. Then,*

$$|\hat{R}_{\text{NN},1}(h) - R_{Q^{\text{test}}}(h)| \geq \min(\mathbb{E}[Y^{test}], 1 - \mathbb{E}[Y^{test}]). \tag{45}$$

In particular, if $\mathbb{E}[Y^{test}] = 1/2$, then one-nearest neighbor risk estimation has error $1/2$.

Proof. Because $S_m^{\text{val}} \overset{\text{iid}}{\sim} \mu$, and μ has Lebesgue density, the nearest neighbor to S is almost surely unique. This implies that $\hat{R}_{\text{NN},1}(h) \in \{0, 1\}$. Therefore,

$$|\hat{R}_{\text{NN},1}(h) - R_{Q^{\text{test}}}(h)| \geq \min_{a \in \{0,1\}} |a - \mathbb{E}[Y^{\text{test}}]| = \min(\mathbb{E}[Y^{\text{test}}], 1 - \mathbb{E}[Y^{\text{test}}]). \tag{46}$$

□

D.3.5. Inconsistency of Nearest Neighbors with Number of Neighbors Depending On Number of Validation Points

We now restate and prove Prop. 4.4, which states that nearest-neighbor risk estimation with the number of neighbors depending (only) on the number of validation points is inconsistent under infill asymptotics, regardless of type of dependence.

Proposition 4.4 (Inconsistency of kNN depending on number of validation points)**.** *Let $(k_n)_{n=1}^{\infty}$ be any sequence of natural numbers. Define the sequence of estimators $\hat{R}_{N^{\text{val}}}$ to be the nearest neighbor risk estimators using N^{val} validation points and $k_{N^{\text{val}}}$ neighbors. Then there exists a data-generating process satisfying infill asymptotics, a test set containing a single point, a predictive method h resulting in an error function satisfying the Lipschitz assumption, and an $\epsilon, \delta > 0$ such that with probability at least $1-\delta$, $\forall N^{\text{val}}$, $|\hat{R}_{N^{\text{val}}}(h) - R_{Q^{\text{test}}}(h)| \geq \epsilon$.*

Proof. The idea is that either 1.) $(k_n)_{n=1}^{\infty}$ has a bounded sub-sequence, in which case along this sub-sequence, the $\hat{R}_{N^{\text{val}}}(h)$ can have a variance bounded below by 0, and so by Prop. D.8 these estimators are bounded away from the test risk with fixed probability. Or 2.) the number of neighbors used tends to infinity, in which case we can find a sequence of data that accumulates more slowly around the test site, leading to many neighbors far from the point being used in the estimator, and therefore non-negligible bias.

We split into these two cases, and give an example showing in either case $\hat{R}_{N^{\text{val}}}$ can be inconsistent.

Case 1: $\liminf_{n\to\infty} k_n < \infty$**.**

Consider a data-generating process with no covariates, $S^{\text{val}} \sim U(0,1)$, $S^{\text{test}} = \{\frac{1}{2}\}$, $Y_n | S_n = \epsilon_n \sim U(-1/2, 1/2)$, $h = 0$ and $\ell(y, y') = |y - y'|$. Because $\liminf_{n\to\infty} k_n < \infty$, there exists a $C > 0$ such that $(k_n)_{n=1}^{\infty}$ contains a bounded sub-sequence $(\tilde{k}_n)_{n=1}^{\infty}$ with $\tilde{k}_n \leq C$ for all $C > 0$. Since the limit superior of a sub-sequence cannot be larger than of the full sequence

$$\lim \sup_{N^{\text{val}}\to\infty} \text{Var}(\hat{R}_{N^{\text{val}}}(h)) \geq \lim \sup_{n\to\infty} \text{Var}(\tilde{R}_n(h)). \tag{47}$$

where $\tilde{R}_n$ denotes the sub-sequence of $\hat{R}_{N^{\text{val}}}$ where the number of validation points runs along the sub-sequence corresponding to $(\tilde{k}_n)_{n=1}^{\infty}$.

Almost surely, for any N^{val} the test point $1/2$ has exactly $k_{N^{\text{val}}}$-nearest neighbors, because the probability that two validation points are equidistant from $1/2$ is 0. Since a countable union of almost sure events is also an almost sure event, with probability 1 for all N^{val} the test point at $1/2$ has exactly $k_{N^{\text{val}}}$ neighbors. Therefore, with probability 1, for all N^{val} the vector of weights $w^{\text{NN},k}$ has exactly $k_{N^{\text{val}}}$ non-zero entries, each with value $1/k_{N^{\text{val}}}$. We condition on this probability 1 event moving forward. In this case,

$$\text{Var}(\hat{R}_{N^{\text{val}}}) = \text{Var}(\sum_{n=1}^{N^{\text{val}}} w^{\text{NN},k} |\epsilon_n|) \tag{48}$$

$$= \frac{1}{k_{N^{\text{val}}}^2} \sum_{i=1}^{k_{N^{\text{val}}}} \text{Var}(|\tilde{\epsilon}_i|) \tag{49}$$

$$= \frac{1}{48 k_{N^{\text{val}}}} \tag{50}$$

where $(\tilde{\epsilon})_{i=1}^{k_{N^{\text{val}}}}$ are the subset of $(\epsilon_n)_{n=1}^{N^{\text{val}}}$ corresponding to the $k_{N^{\text{val}}}$ nearest neighbors to $1/2$. The factor of 48 comes from the variance of a uniform random variable on $[0, 1/2]$.

For all N^{val} corresponding to the bounded sub-sequence $(\tilde{k}_n)_{n=1}^{\infty}$, we conclude

$$\text{Var}(\tilde{R}_n(h)) \geq \frac{1}{48C}. \tag{51}$$

Therefore,

$$\lim \sup_{N^{\text{val}}\to\infty} \text{Var}(\hat{R}_{N^{\text{val}}}(h)) \geq \frac{1}{48C}. \tag{52}$$

Applying Prop. D.8 to the random variable $|\hat{R}_{N^{\text{val}}}(h) - R(h)|$ we conclude there exists an ϵ, δ such that with probability at least $1 - \delta$,

$$\lim \sup_{N^{\text{val}}\to\infty} |\hat{R}_{N^{\text{val}}}(h) - R(h)| \geq \epsilon. \tag{53}$$

Case 2: $\liminf_{n\to\infty} k_n = \infty$**.**

Consider the data generating process, $Y_n|S_n = S_n$ on $[0,1]$ with $S^{\text{test}} = \{0\}$ and $\ell(y,y') = |y-y'|$ and $h = 0$. We then have $R(h) = 0$. We will construct a sequence of validation sites S^{val} such that for each N^{val} less than $k_{N^{\text{val}}}/2$ of the validation sites fall in the interval $[0, 1/4)$. For any such sequence (supposing such a sequence exists for the moment),

$$|\hat{R}_{N^{\text{val}}}(h) - R(h)| = |\hat{R}_{N^{\text{val}}}(h)| \geq \frac{1}{k_{N^{\text{val}}}} \cdot \frac{k_{N^{\text{val}}}}{2} \frac{1}{4} \geq \frac{1}{8}. \tag{54}$$

All that remains is to construct such a sequence that also satisfies infill asymptotics. Define the function $\psi : \mathbb{N} \to (0,1)$ by

$$\psi(i) = \frac{i - 2^{\lfloor \log_2 i \rfloor + 1}}{2^{\lfloor \log_2 i \rfloor + 1}}. \tag{55}$$

This corresponds to the dyadic sequence $(1/2, 1/4, 3/4, 1/8, 3/8, 5/8, 7/8, 1/16, \dots)$. The essential properties of this function for our application is that the image of the function is dense on $(0,1)$ and for any i at least $i/2$ of $(\phi(j))_{j=1}^{i}$ are in the interval $(0, 1/2]$.

Algorithm 1 Algorithm defining $(S_n^{\text{val}})_{n=1}^{\infty}$

while True: **do**
  if $j < k_n$ and $n - j > j$ **then**
    $S_n^{\text{val}} = \frac{1}{2} - \frac{1}{2}\psi(j)$
    $j \leftarrow j + 1$.
  else
    $S_n^{\text{val}} = \frac{1}{2} + \frac{1}{2}\phi(n-j)$
    $n \leftarrow n + 1$.
  end if
end while

The validation points are defined algorithmically via Algorithm 1. Because k_n is unbounded, the first condition must be called infinitely often, and so j eventually takes on all natural numbers in this loop.

Because $n - j > j$ each time the first condition is called, $n - j$ is incremented if and only if j is not incremented $n - j$ also takes on all natural numbers in this loop. Therefore, $(S_n^{\text{val}})_{n=1}^{\infty}$ is a dense set in $[0,1]$ because $\frac{1}{2} - \frac{1}{2}\psi(\mathbb{N})$ is dense on $[0, 1/2]$ and $\frac{1}{2} + \frac{1}{2}\psi(\mathbb{N})$ is dense on $[1/2, 1]$. We conclude this sequence satisfies the infill asymptotics.

For any N^{val}, the number of validation points less than $1/2$ is not more than $k_{N^{\text{val}}}$ by induction and using the first condition in the if statement. Of the points placed in $(0, 1/2)$ at most half of them are in $(0, 1/4]$, by our earlier observation that for any i at least $i/2$ of $(\psi(j))_{j=1}^{i}$ are in the interval $(0, 1/2]$. Therefore, not more than $k_{N^{\text{val}}}/2$ points are in $[0, 1/4)$ for any N^{val}. □

D.4. General Nearest Neighbor Bound and Selecting the Number of Neighbors

In this section, we prove two results giving bounds on the performance of nearest neighbor risk estimation that will be the basis of later results. The first, already stated in the main text, is a general bound for any k.

Theorem 5.1 (Bound on Estimation Error in Terms of Fill Distance)**.** *Consider a validation set D^{val} of size N^{val} and a test set D^{test} of size M^{test}. Take the k-nearest neighbors test-risk estimator from Def. 4.2. Choose $\delta \in (0,1)$ and k such that $1 \leq k \leq N^{\text{val}}$. Let $\rho_k := \zeta^k(S^{\text{val}}_{1:N^{\text{val}}}, S^{\text{test}}_{1:M^{\text{test}}})$ and $\beta_\delta := \Delta\sqrt{\frac{1}{2}\log\frac{2}{\delta}}$. Take Assumptions 2.1, 2.2, 2.4 and 2.5. Then, with probability at least $1-\delta$,*

$$|R_{Q^{\text{test}}}(h) - \hat{R}_{\text{NN},k}(h)| \ \leq L\rho_k + \beta_\delta \|w^{\text{NN},k}\|_2 \tag{4}$$

$$\leq L\rho_k + \beta_\delta \sqrt{\max_{1\leq n\leq N^{\text{val}}} Q^{\text{test}}(B(S_n^{\text{val}}, \rho_k))/k}, \tag{5}$$

where $B(S,r)$ denotes the ball of radius r centered at S. See Assumptions 2.1 and 2.5 and Def. 2.3 for Δ, L, Q^{test} respectively.

The second result we will show is specific to SNN and relates the error incurred by using $k^\star_{T_2}$ to the error of the minimizer of the bound from Thm. 5.1:

Proposition D.10 (Minimization over Powers of Two)**.** *Let $T = \{1, \dots, N^{\text{val}}\}$ and $T_2 = \{2^i\}_{i=1}^{\lfloor \log_2 N^{\text{val}} \rfloor}$. Fix $\delta \in (0,1)$. Define $k^\star_{T_2} \in \arg\min_{k \in T_2} \rho_k + \beta_\delta \|w^{\text{NN},k}\|_2$ with $\beta_\delta = \Delta\sqrt{\frac{1}{2}\log\frac{2}{\delta}}$. Define $C_L = \max(1, L)$. Under Assumptions 2.1, 2.2 and 2.5 with probability at least $1-\delta$*

$$|R_{Q^{\text{test}}}(h) - \hat{R}_{\text{NN},k^\star_{T_2}}(h)| \leq \sqrt{2} C_L \left(\min_{k \in T} \rho_k + \beta_\delta \sqrt{\max_{1 \leq n \leq N^{\text{val}}} \frac{Q^{\text{test}}(B(S_n^{\text{val}}, \rho_k))}{k}} \right), \tag{56}$$

where $B(S, r)$ denotes the ball of radius r centered at S.

D.4.1. Preliminary Result: Hoeffding's Inequality

We begin by recalling Hoeffding's inequality, which will be used to control tail probabilities of the sum of the weighted losses being far from its expectation.

Lemma D.11 (Hoeffding's Inequality, (Hoeffding, 1963, Theorem 2))**.** *Let $(Z_i)_{i=1}^\ell$ be independent random variables and $(a_i)_{i=1}^\ell$ and $(b_i)_{i=1}^\ell$ be sequences of real numbers such that $a_i \leq Z_i \leq b_i$ almost surely. Then for all $t > 0$*

$$\Pr\left(\left| \sum_{i=1}^\ell Z_i - \sum_{i=1}^\ell \mathbb{E}[Z_i] \right| \geq t \right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^\ell (b_i - a_i)^2} \right). \tag{57}$$

Equivalently, for any $\delta \in (0,1)$ with probability at least $1-\delta$

$$\left| \sum_{i=1}^\ell Z_i - \sum_{i=1}^\ell \mathbb{E}[Z_i] \right| \leq \|b - a\|_2 \sqrt{\frac{1}{2}\log\frac{2}{\delta}}. \tag{58}$$

where $a, b \in \mathbb{R}^\ell$ have entries a_i and b_i respectively.

D.4.2. Proof of General Nearest Neighbor Risk Estimation Bound

We again recall and prove Thm. 5.1. The idea is to apply triangle inequality to split the error into a bias term and a sampling error term. The sampling error term is then controlled with Lemma D.11 since the loss is bounded. The bias term is controlled using Assumption 2.5.

Theorem 5.1 (Bound on Estimation Error in Terms of Fill Distance)**.** *Consider a validation set D^{val} of size N^{val} and a test set D^{test} of size M^{test}. Take the k-nearest neighbors test-risk estimator from Def. 4.2. Choose $\delta \in (0,1)$ and k such that $1 \leq k \leq N^{\text{val}}$. Let $\rho_k := \zeta^k(S^{\text{val}}_{1:N^{\text{val}}}, S^{\text{test}}_{1:M^{\text{test}}})$ and $\beta_\delta := \Delta\sqrt{\frac{1}{2}\log\frac{2}{\delta}}$. Take Assumptions 2.1, 2.2, 2.4 and 2.5. Then, with probability at least $1-\delta$,*

$$|R_{Q^{\text{test}}}(h) - \hat{R}_{\text{NN},k}(h)| \leq L\rho_k + \beta_\delta \|w^{\text{NN},k}\|_2 \tag{4}$$

$$\leq L\rho_k + \beta_\delta \sqrt{\max_{1 \leq n \leq N^{\text{val}}} Q^{\text{test}}(B(S_n^{\text{val}}, \rho_k))/k}, \tag{5}$$

where $B(S, r)$ denotes the ball of radius r centered at S. See Assumptions 2.1 and 2.5 and Def. 2.3 for Δ, L, Q^{test} respectively.

Proof. By the triangle inequality,

$$|R_{Q^{\text{test}}}(h) - \hat{R}_{\text{NN},k}(h)| \leq \underbrace{\left| R_{Q^{\text{test}}}(h) - \sum_{n=1}^{N^{\text{val}}} w_n^{\text{NN},k} e_h(S_n^{\text{val}}) \right|}_{\tau_1} \tag{59}$$

$$+ \underbrace{\left| \sum_{n=1}^{N^{\text{val}}} w_n^{\text{NN},k} (\ell(f(S_n^{\text{val}}, \chi(S_n^{\text{val}}), \epsilon_n^{\text{val}}), h^\chi(S_n^{\text{val}})) - e_h(S_n^{\text{val}}) \right|}_{\tau_2}. \tag{60}$$

The first term, τ_1 is a bias term, while the second term, τ_2 is a sum of N^{val} independent variables with expectation zero. Using Assumption 2.1, we apply Hoeffding's inequality (Lemma D.11) to bound τ_2: for any $\delta \in (0,1)$ with probability at least $1-\delta$

$$\tau_2 \leq \Delta \|w^{\text{NN},k}\|_2 \sqrt{\frac{1}{2}\log\frac{2}{\delta}}. \tag{61}$$

By Hölder's inequality and because the weights are non-negative and sum to one,

$$\|w^{\text{NN},k}\|_2 \leq \sqrt{\|w^{\text{NN},k}\|_1 \|w^{\text{NN},k}\|_\infty} = \sqrt{\max_{1\leq n\leq N^{\text{val}}} w_n^{\text{NN},k}}. \tag{62}$$

Recalling the definition of $w^{\text{NN},k}$ (Def. 4.2) and that each $A^k(s)$ contains at least k points by construction,

$$w_n^{\text{NN},k} = \frac{1}{M^{\text{test}}} \sum_{m=1}^{M^{\text{test}}} \frac{1}{|A^k(S_m^{\text{test}})|} \mathbf{1}\{S_n^{\text{val}} \in A^k(S_m^{\text{test}})\} \leq \frac{1}{k} \frac{1}{M^{\text{test}}} \sum_{m=1}^{M^{\text{test}}} \mathbf{1}\{S_n^{\text{val}} \in A^k(S_m^{\text{test}})\}. \tag{63}$$

By the definition of ρ_k,

$$S_n^{\text{val}} \in A^k(S_m^{\text{test}}) \Rightarrow S_m^{\text{test}} \in B(S_n^{\text{val}}, \rho_k) \tag{64}$$

and so

$$\mathbf{1}\{S_n^{\text{val}} \in A^k(S_m^{\text{test}})\} \leq \mathbf{1}\{S_m^{\text{test}} \in B(S_n^{\text{val}}, \rho_k)\}. \tag{65}$$

Substituting this in Eqn. (63) and using Eqn. (62)

$$\|w^{\text{NN},k}\|_2 \leq \sqrt{\frac{1}{k} \max_{1\leq n\leq N^{\text{val}}} \frac{1}{M^{\text{test}}} \sum_{m=1}^{M^{\text{test}}} \mathbf{1}\{S_m^{\text{test}} \in B(S_n^{\text{val}}, \rho_k)\}} = \sqrt{\max_{1\leq n\leq N^{\text{val}}} \frac{Q^{\text{test}}(B(S_n^{\text{val}}, \rho_k))}{k}}. \tag{66}$$

It remains to bound the bias term, τ_1. Define $\alpha_{nm}^k = \frac{1}{|A^k(S_m^{\text{test}})|}\mathbf{1}\{S_n^{\text{val}} \in A^k(S_m^{\text{test}})\}$. Recalling the definition of $w^{\text{NN},k}$ and rearranging the order of summation

$$\tau_1 = \left| \frac{1}{M^{\text{test}}} \sum_{m=1}^{M^{\text{test}}} (e_h(S_m^{\text{test}}) - \sum_{n=1}^{N^{\text{val}}} \alpha_{nm}^k e_h(S_n^{\text{val}})) \right| \tag{67}$$

$$\leq \max_{1\leq m\leq M^{\text{test}}} \left| e_h(S_m^{\text{test}}) - \sum_{n=1}^{N^{\text{val}}} \alpha_{nm}^k e_h(S_n^{\text{val}}) \right|. \tag{68}$$

Because for any $1 \leq m \leq M^{\text{test}}$, $\sum_{n=1}^{N^{\text{val}}} \alpha_{nm}^k = 1$

$$\left| e_h(S_m^{\text{test}}) - \sum_{n=1}^{N^{\text{val}}} \alpha_{mn}^k e_h(S_n^{\text{val}}) \right| = \left| \sum_{n=1}^{N^{\text{val}}} \alpha_{mn}^k (e_h(S_m^{\text{test}}) - e_h(S_n^{\text{val}})) \right| \tag{69}$$

$$\leq \max_{n:\alpha_{mn}^k>0} |e_h(S_m^{\text{test}}) - e_h(S_n^{\text{val}})|. \tag{70}$$

Applying Assumption 2.5 and taking the maximum over m as well,

$$\tau_1 \leq \max_{n,m:\alpha_{nm}^k>0} L d_{\mathcal{S}}(S_m^{\text{test}}, S_n^{\text{val}}). \tag{71}$$

The constraint $\alpha_{nm}^k > 0$ implies that $d_{\mathcal{S}}(S_m^{\text{test}}, S_n^{\text{val}}) \leq \rho_k$ and so

$$\tau_1 \leq L\rho_k. \tag{72}$$

The result follows from combining Eqns. (60), (61), (66) and (72). □

D.4.3. Proofs Related to Selecting the Number of Neighbors

We now restate and prove a bound that upper bounds the error of risk estimation with SNN (i.e. using $k^\star_{T_2}$ neighbors) to the minimum of the upper bound from Thm. 5.1 over all k. A key observation is that because Thm. 5.1 is conditional on the test locations, it can be minimized without the need to take a union bound over all k in the set we minimize over. We first need a preliminary result, which holds for minimization over any subset of $\{1, \ldots, N^{\text{val}}\}$.

Proposition D.12 (Minimization of Upper Bound)**.** *Let $T \subset \{1, \ldots, N^{\text{val}}\}$. Fix $\delta \in (0,1)$. Define $k^\star_T \in \arg\min_{k \in T} \rho_k + \beta_\delta \|w^{\text{NN},k}\|_2$ with $\beta_\delta = \Delta\sqrt{\frac{1}{2}\log\frac{2}{\delta}}$. Define $C_L = \max(1, L)$. Under Assumptions 2.1, 2.2 and 2.5 with probability at least $1-\delta$*

$$|R_{Q^{\text{test}}}(h) - \hat{R}_{\text{NN},k^\star_T}(h)| \leq C_L \left(\min_{k \in T} \rho_k + \beta_\delta \|w^{\text{NN},k}\|_2 \right) \tag{73}$$

$$\leq C_L \left(\min_{k \in T} \rho_k + \beta_\delta \sqrt{\max_{1 \leq n \leq N^{\text{val}}} \frac{Q^{\text{test}}(B(S^{\text{val}}_n, \rho_k))}{k}} \right), \tag{74}$$

where $B(S, r)$ denotes the ball of radius r centered at S.

Proof of Prop. D.12. Because the minimization problem does not depend on a quantity that is treated as random in Thm. 5.1, we may directly apply Thm. 5.1 to $k^\star_T$ to conclude with probability at least $1-\delta$

$$|R(h) - \hat{R}_{\text{NN},k^\star_T}(h)| \leq L\rho_{k^\star_T} + \Delta\|w^{\text{NN},k^\star_T}\|_2\sqrt{\frac{1}{2}\log\frac{2}{\delta}} \tag{75}$$

We split into cases.

Case 1: $L \leq 1$ Because $L \leq 1$ and by the minimality of $k^\star_T$,

$$L\rho_{k^\star_T} + \Delta\|w^{\text{NN},k^\star_T}\|_2\sqrt{\frac{1}{2}\log\frac{2}{\delta}} \leq \rho_{k^\star_T} + \Delta\|w^{\text{NN},k^\star_T}\|_2\sqrt{\frac{1}{2}\log\frac{2}{\delta}} \tag{76}$$

$$= \min_{k \in T} \rho_k + \Delta\|w^{\text{NN},k}\|_2\sqrt{\frac{1}{2}\log\frac{2}{\delta}}. \tag{77}$$

Case 2: $L > 1$ Because $L > 1$ and the second term is non-negative,

$$L\rho_{k^\star_T} + \Delta\|w^{\text{NN},k^\star_T}\|_2\sqrt{\frac{1}{2}\log\frac{2}{\delta}} \leq L\left(\rho_{k^\star_T} + \Delta\|w^{\text{NN},k^\star_T}\|_2\sqrt{\frac{1}{2}\log\frac{2}{\delta}}\right) \tag{78}$$

$$= L\min_{k \in T} \rho_k + \Delta\|w^{\text{NN},k}\|_2\sqrt{\frac{1}{2}\log\frac{2}{\delta}}. \tag{79}$$

Combining Eqns. (75), (77) and (79) gives the result. □

Proposition D.10 (Minimization over Powers of Two)**.** *Let $T = \{1, \ldots, N^{\text{val}}\}$ and $T_2 = \{2^i\}_{i=1}^{\lfloor \log_2 N^{\text{val}} \rfloor}$. Fix $\delta \in (0,1)$. Define $k^\star_{T_2} \in \arg\min_{k \in T_2} \rho_k + \beta_\delta \|w^{\text{NN},k}\|_2$ with $\beta_\delta = \Delta\sqrt{\frac{1}{2}\log\frac{2}{\delta}}$. Define $C_L = \max(1, L)$. Under Assumptions 2.1, 2.2 and 2.5 with probability at least $1-\delta$*

$$|R_{Q^{\text{test}}}(h) - \hat{R}_{\text{NN},k^\star_{T_2}}(h)| \leq \sqrt{2}C_L \left(\min_{k \in T} \rho_k + \beta_\delta \sqrt{\max_{1 \leq n \leq N^{\text{val}}} \frac{Q^{\text{test}}(B(S^{\text{val}}_n, \rho_k))}{k}} \right), \tag{56}$$

where $B(S, r)$ denotes the ball of radius r centered at S.

Proof. Let $k^\star_T$ denote a minimizer of the bound on the right hand side, which exists since the minimization is over a finite set. If $k^\star_T = 1$, we are done since $1 \in T_2$. Otherwise, there exists a $\tilde{k} \in T_2$ such that,

$$k^\star_T/2 \leq \tilde{k} \leq k^\star_T. \tag{80}$$

By monotonicity of the k^{th} order fill distance in k,

$$\rho_{\tilde{k}} \leq \rho_{k_T^\star}. \tag{81}$$

This also implies,

$$Q^{\text{test}}(B(S_n^{\text{val}}, \rho_{\tilde{k}})) \leq Q^{\text{test}}(B(S_n^{\text{val}}, \rho_{k_T^\star})), \tag{82}$$

since the measure of a subset is never larger than the measure of a set that contains it. Therefore,

$$\min_{k \in T_2} \left(\rho_k + \beta_\delta \sqrt{\max_{1 \leq n \leq N^{\text{val}}} \frac{Q^{\text{test}}(B(S_n^{\text{val}}, \rho_k))}{k}} \right) \leq \rho_{\tilde{k}} + \beta_\delta \sqrt{\max_{1 \leq n \leq N^{\text{val}}} \frac{Q^{\text{test}}(B(S_n^{\text{val}}, \rho_{\tilde{k}}))}{\tilde{k}}} \tag{83}$$

$$\leq \rho_{k_T^\star} + \beta_\delta \sqrt{2} \sqrt{\max_{1 \leq n \leq N^{\text{val}}} \frac{Q^{\text{test}}(B(S_n^{\text{val}}, \rho_{k_T^\star}))}{k_T^\star}} \tag{84}$$

$$\leq \sqrt{2} (\rho_{k_T^\star} + \beta_\delta \sqrt{\max_{1 \leq n \leq N^{\text{val}}} \frac{Q^{\text{test}}(B(S_n^{\text{val}}, \rho_{k_T^\star}))}{k_T^\star}}) \tag{85}$$

$$= \sqrt{2} \left(\min_{k \in T} \rho_k + \beta_\delta \sqrt{\max_{1 \leq n \leq N^{\text{val}}} \frac{Q^{\text{test}}(B(S_n^{\text{val}}, \rho_k))}{k}} \right). \tag{86}$$

The result now follows from Prop. D.12. □

D.5. Consistency of our Nearest Neighbor Method under Infill Asymptotics

In this section we prove Cor. 5.2, which establishes the consistency of SNN under infill asymptotics. The idea of the proof is to upper bound the k^{th} order fill distance of the validation points in the test points to the fill distance of the validation points in $[0,1]^d$. This allows together with Thm. 5.1 and Prop. D.10 allows us to derive an upper bound on the error of our method in terms of the fill distance, from which the result follows.

D.5.1. Preliminary Lemma: Relating Fill Distances

Proposition D.13. *Let A be an ϵ-net for $[0,1]^d$. Then for $k \leq \left(\frac{1}{2\epsilon}\right)^d$*

$$\zeta^k(A, [0,1]^d) \leq 2k^{1/d}\epsilon + \epsilon. \tag{87}$$

Proof. Let $S \in [0,1]^d$ and $\tau \in (0,1]$. Then,

$$|A \cap B(S, \tau + \epsilon)| \geq \mathfrak{N}(\epsilon, B(S, \tau)) \tag{88}$$

because for any $S' \in B(S,\tau) \cap [0,1]^d$, there is a point $a \in A$ such that $d(S', a) \leq \epsilon$ and for any such point, this a must also be in $B(s, \tau + \epsilon)$ by the triangle inequality. Let C be an ϵ-net of $B(s,\tau) \cap [0,1]^d$. Then by the definition of a net and by subadditivity

$$\text{vol}(B(S,\tau) \cap [0,1]^d) \leq \text{vol}(\cup_{c \in C} B(c, \epsilon)) \leq |C| B_d \frac{1}{\epsilon^d} \tag{89}$$

Since $S \in [0,1]^d$ and $\tau \leq 1$, at least one orthant of $B(s,\tau)$ is contained in $[0,1]^d$, so

$$\text{vol}(B(S,\tau) \cap [0,1]^d) \geq 2^{-d} B_d \frac{1}{\tau^d}. \tag{90}$$

Combining the previous estimates,

$$\mathfrak{N}(\epsilon, B(S,\tau) \cap [0,1]^d) \geq \left(\frac{\tau}{2\epsilon}\right)^d \tag{91}$$

Choose $\tau = 2\epsilon k^{1/d} \in (0,1)$. Then

$$|A \cap B(S, \tau + \epsilon)| \geq k. \tag{92}$$

As S was arbitrary, this holds for all $S \in [0,1]^d$, and so for all $S \in [0,1]^d$, there are k points in A in $B(S, 2\epsilon k^{1/d} + \epsilon)$. □

D.5.2. Bound on Loss Depending on Fill Distance

We now present and prove an upper bound on the error of SNN that depends on the fill distance of the validation set in $[0,1]^d$. We will derive consistency of the estimator under infill asymptotics as a corollary of this bound. The bound will also be relevant in later discussion of model selection, where we are also interested in rates of convergence of estimators.

Corollary D.14 (Bound for Dense Validation Data and General Test Data)**.** *Suppose that $\mathcal{S} = [0,1]^d$ and Assumptions 2.1, 2.2 and 2.5. Let $k^\star_{T_2} \in \arg\min_{k \in T_2} \rho_k + \beta_\delta \|w^{\mathrm{NN},k}\|_2$ with $\beta_\delta = \Delta\sqrt{\frac{1}{2}\log\frac{2}{\delta}}$. Then there exists a constant $K_{d,\delta,\Delta,L}$ such that with probability at least $1-\delta$*

$$|R_{Q^{\text{test}}}(h) - \hat{R}_{\mathrm{NN},k^\star_{T_2}}(h)| \leq K_{d,\delta,\Delta,L}\tilde{\rho}^{\frac{d}{d+2}}, \tag{93}$$

where $\tilde{\rho} = \zeta(S^{\text{val}}_{1:N^{\text{val}}}, [0,1]^d)$.

Proof. For all, $\tilde{\rho} \geq 1$, the stated bound holds with $K = L\sqrt{d} + \beta_\delta$. Therefore, moving forward, we assume $\tilde{\rho} < 1$.

Choose $k = \min(\lceil(\gamma\tilde{\rho})^{-\frac{2d}{d+2}}\rceil, N^{\text{val}})$ for some $\gamma \in (0, 1/2)$ to be specified later. Because $\frac{2d}{d+2} < d$ and $\gamma \leq \frac{1}{2}$, this k satisfies the conditions of Prop. D.13 so,

$$\begin{aligned}
\rho_k \leq \zeta^k(A, [0,1]^d) &\leq 2\min(\lceil(\gamma\tilde{\rho})^{-\frac{2d}{d+2}}\rceil^{1/d}, (N^{\text{val}})^{1/d})\tilde{\rho} + \tilde{\rho} && (94)\\
&\leq 2(1+\gamma\tilde{\rho})^{-\frac{2}{d+2}})\tilde{\rho} + \tilde{\rho} && (95)\\
&\leq 4\gamma^{-2}\tilde{\rho}^{\frac{d}{d+2}}. && (96)
\end{aligned}$$

We now apply Thm. 5.1 and Prop. D.10 together with this bound to conclude with probability $1-\delta$

$$\begin{aligned}
|R_{Q^{\text{test}}}(h) - \hat{R}_{\mathrm{NN},k^\star_{T_2}}(h)| &\leq \sqrt{2}C_L(4\gamma^{-2}\tilde{\rho}^{\frac{d}{d+2}} + \frac{1}{\sqrt{k}}\beta_\delta) && (97)\\
&\leq \sqrt{2}C_L\left(4\gamma^{-2}\tilde{\rho}^{\frac{d}{d+2}} + \min\left(\gamma^{\frac{d}{d+2}}\tilde{\rho}^{\frac{d}{d+2}}, \frac{1}{\sqrt{N^{\text{val}}}}\right)\beta_\delta\right). && (98)
\end{aligned}$$

Choosing $\gamma = \frac{1}{4}$ (for example) completes the proof. □

D.5.3. Consistency of Spatial Nearest Neighbors

We now restate and prove that the spatial nearest neighbor procedure we describe is consistent under infill asymptotics. This follows as a corollary of Cor. D.14 since the infill assumption means that for any fixed δ, the upper bound in Cor. D.14 tends to zero with the fill distance.

Corollary 5.2 (Our Method is Consistent under Infill Asymptotics)**.** *Let $\mathcal{S} = [0,1]^d$. Take Assumptions 2.1, 2.2, 2.4 and 2.5. Let $\tilde{\rho} := \zeta(S^{\text{val}}_{1:N}, \mathcal{S})$. Let $k^\star_{T_2} \in \arg\min_{k \in T_2} \rho_k + \beta_\delta \|w^{\mathrm{NN},k}\|_2$ with $\delta = \min(1, r)$ and $r \in [c\tilde{\rho}, C\tilde{\rho}]$ for some constants (possibly depending on dimension) $c, C > 0$. Then the $k^\star_{T_2}$-nearest neighbor risk estimator is consistent under infill asymptotics.*

Proof. Take $\delta = \min(1, r)$. Under infill asymptotics, this tends to zero because $r \leq C\tilde{\rho}$ and $\tilde{\rho}$ tends to 0 by assumption. Cor. D.14 (or more precisely Eqn. (98) which makes the depend of the bound on δ explicit), with probability at least $1-\delta$

$$\begin{aligned}
|R_{Q^{\text{test}}}(h) - \hat{R}_{\mathrm{NN},k^\star_{T_2}}(h)| &\leq K_{d,\Delta,L}\tilde{\rho}^{d/(d+2)}\sqrt{\log\frac{1}{r}} && (99)\\
&\leq K_{d,\Delta,L}\tilde{\rho}^{d/(d+2)}\sqrt{\log\frac{1}{c\tilde{\rho}}}. && (100)
\end{aligned}$$

The left hand side of this bound tends to zero with $\tilde{\rho}$, and so $\hat{R}_{\mathrm{NN},k^\star_{T_2}}(h)$ converges in probability to $R_{Q^{\text{test}}}(h)$. □

D.5.4. Computation of An Approximate Fill Distance

The fill distance can be computed exactly by computing the vertices of a d-dimensional Voronoi diagram with the validation points, then computing the maximum distance from each of these vertices to a point in the validation set. For problems in 1 or 2 dimensions, this is feasible even with a large number of validation points, because algorithms for computation of the Voronoi diagram can be done in nearly linear, $O(N^{\text{val}} \log N^{\text{val}})$ time in 1 and 2 dimensions (Okabe et al., 2000, Chapter 4). In higher dimensions, the worst case complexity of algorithms for computing Voronoi diagrams can be $O((N^{\text{val}})^{\lfloor d/2 \rfloor} + N^{\text{val}} \log N^{\text{val}})$, and so for large number of points in more the computational cost can become quite high.

To avoid this computational cost, we instead use a simple space partitioning algorithm when $\mathcal{S} = [0,1]^d$ that is guaranteed to give an approximation to the fill distance r satisfying $\frac{\tilde{\rho}}{2\sqrt{d}} \leq r \leq 2\tilde{\rho}$. The idea is to split the domain into 2^d quadrants and check if each quadrant contains a validation point. If it does, recurse, otherwise stop and keep track of the side length of each quadrant, call it r. When the algorithm terminates it must be the case that there exists a partitioning of $[0,1]^d$ into cubes of side length $2r$, with each cube containing at least one validation point. Because this is a partition, each point in the spatial domain must also be within a cube, and so the fill distance is upper bounded by the maximum distance between two points in a cube of side length $2r$, i.e. $\tilde{\rho} \leq 2r\sqrt{d}$. Also, when the algorithm stops, a cube of side length r has been found that does not contain any points. Therefore the fill distance is lower bounded by the distance from the center of this cube to the closest validation point, which must be at least $r/2$. That is, $\tilde{\rho} \geq r/2$. Rearranging, we see that

$$\frac{\tilde{\rho}}{2\sqrt{d}} \leq r \leq 2\tilde{\rho} \tag{101}$$

as claimed. Finally, we address the computational complexity of this approach. The number of times we recurse is $\log_2 r$, which is $O(\log 1/\tilde{\rho})$, which is in turn $O(\log N^{\text{val}})$, because the fill distance cannot decrease faster than the inverse of the covering number of $[0,1]^d$, which certainly does not decrease faster than $1/N^{\text{val}}$.

It remains to consider the complexity of deciding which orthant all of the validation points lie in, and partioning the points by orthant. This can be done by looping over each of the dimensions, partitioning the points in the cube based on whether or not that coordinate is on the left or right hand side of the current cube, which is $O(N^{\text{val}} d)$. Therefore, the total computational complexity of this algorithm is not more than $O(N^{\text{val}} d \log N^{\text{val}})$, which is nearly linear in N^{val}.

D.6. Model Selection: Rates of Convergence of Spatial Nearest Neighbors

We now present an extended version of earlier discussion in App. C, as well as results on rates of convergence of spatial nearest neighbors that provides some support to claims regarding model selection. As a special case, we consider grid prediction (for example for assessing the global performance of a map constructed using a predictive method). We then discuss rates of convergence in the more general setting earlier addressed in Cor. D.14.

D.6.1. A Heuristic Discussion of Model Selection with Increasing Amounts of Data

Suppose we have a fixed test task, and two data-driven algorithms for making predictions. We also suppose we have allocated a fixed percentage of data for training, and the remainder for validation. Can we use spatial nearest neighbors to select between the two method? As the amount of training data increases we would hope that both data-driven algorithms produce better predictive methods. We therefore do not expect consistency to be sufficient to select between the two sequences of predictive methods: if both are converging to the optimal estimator at (possibly different) rates as the amount of training data increases, we want our error in estimating the risk of the two methods to converge to zero faster at a rate faster in the amount of validation data than the rate that the slower converging method converges to the optimal predictor. This would suggest we should be able to reliably identify the better sequence of predictive methods as the amount of data increases.

We will assume we are in the additive, homoskedastic error setting so that $Y = f(S, \chi(S)) + \epsilon$. We will assume that the noise is bounded, and that squared loss is used. Finally, we will assume that training data is also generated following this process. We will also assume that f is Lipschitz continuous. In this setting, minimax pointwise regression rates are $\theta((N^{\text{train}})^{-\frac{1}{d+2}})$ (cf. Tsybakov (2008, Theorem 2.3, Corollary 2.2) in one-dimension under a fixed grid design and Tibshirani (2023, Examples 3.1, 3.2) for a multi-dimensional version with both fixed grid and random design. The latter assumes Gaussian noise instead of bounded noise). Ideally, we would like a method for performing model selection to be able to distinguish between a sequence of predictive methods converging at slower than the minimax optimal rate and a sequence of predictive methods converging at the minimax rate and to be able to reliably select between two sequences of predictive methods both of which are converging at the minimax rate, but with different constants.

We will now present some finite sample and asymptotic bounds on the convergence of spatial nearest neighbors for a fixed hypothesis then return to the question of model selection in light of these results.

D.6.2. Rates of Convergence for Grid Prediction

We first consider the grid prediction task specifically, as this is a common problem in spatial analyses. For example, one might want to reconstruct air temperature across the continental United States on a dense grid (map) based on remotely sensed covariates observed on this grid and sparsely observed weather station data. We consider a test task grid prediction if the data falls on a regular d-dimensional grid.

Assumption D.15 (Grid prediction)**.** We say that a task is *grid prediction* if $\mathcal{S} = [0,1]^d$ and $Q^{\text{test}} = \frac{1}{g^d}\sum_{S\in\{i/g:1\leq i\leq g\}^d}\delta_S$ for some $g \in \mathbb{N}$.

As long as the resolution of the map is high, both 1-nearest neighbor and spatial nearest neighbors provide reliable estimates of the error.

Corollary D.16 (Bound on Estimation Error for Grid Prediction)**.** *With the same assumptions as Thm. 5.1 and additionally Assumption D.15, with probability at least* $1-\delta$

$$|R_{Q^{\text{test}}}(h) - \hat{R}_{\text{NN},1}(h)| \leq L\rho + \beta_\delta\sqrt{\max(\tfrac{2^d}{M^{\text{test}}}, (8\rho)^d)}. \tag{102}$$

Also, with probability at least $1-\delta$

$$|R_{Q^{\text{test}}}(h) - \hat{R}_{\text{NN},k^\star_{T_2}}(h)| \leq C_L\left(\rho + \beta_\delta\sqrt{\max(\tfrac{2^d}{M^{\text{test}}}, (8\rho)^d)}\right).$$

with $\rho = \rho_1, \beta_\delta = \Delta\sqrt{\frac{1}{2}\log\frac{2}{\delta}}$ *and* $C_L = \max(1, L)$.

See App. D.6.5 for a proof. The right-hand side of the bound is small as long as there is a validation point near each test point and the resolution of the map is high. If the available data for validation is generated i.i.d. as in Prop. 3.3, then right hand side becomes, up to logarithmic factors in N^{val}, $\frac{1}{\sqrt{M^{\text{test}}}} + (N^{\text{val}})^{\min(-1/2,-1/d)}$

D.6.3. Statement and Discussion of Result for IID validation data and Grid Prediction

Corollary D.17 (Convergence of Grid Prediction with Independent and Identically Distributed Validation Data)**.** *Suppose that* $\mathcal{S} = [0,1]^d$, $S_n^{\text{val}} \overset{\text{iid}}{\sim} P$ *for* $1 \leq n \leq N^{\text{val}}$ *with* $N^{\text{val}} > 1$ *and* P *has Lebesgue density lower bounded by* $c > 0$. *Additionally, take the assumptions of Cor. D.16. Fix* $\delta \in (0,1)$ *and* $k \in \{1, k^\star_{T_2}\}$. *Then there exists a constant* $K_{d,\delta,L,\Delta,c}$ *that depends only on* d, δ, L, c *and* Δ *such that with probability at least* $1-\delta$

$$|\hat{R}_{\text{NN},k}(h) - R_{Q^{\text{test}}}(h)| \leq K_{d,\delta,L,\Delta,c}\left((\tfrac{\log N^{\text{val}}}{N^{\text{val}}})^{\min(\frac{1}{2},\frac{1}{d})} + \tfrac{1}{\sqrt{M^{\text{test}}}}\right).$$

Proof. From Prop. 3.3, there exists a constant γ depending on d, L, δ, c, Δ such that with probability at least $1-\delta/2$

$$\rho \leq \gamma\left(\frac{\log N^{\text{val}}}{N^{\text{val}}}\right)^{1/d}. \tag{103}$$

An upper bound on the bound in Cor. D.16 shows that for $k \in \{1, k^\star_{T_2}\}$ with probability $1-\delta/2$,

$$|\hat{R}_{\text{NN},k}(h) - R_{Q^{\text{test}}}(h)| \leq \gamma C_L \max(\beta_{\delta/2}8^{d/2}, 1)\left(\rho + \frac{1}{\sqrt{M^{\text{test}}}} + \rho^{d/2}\right). \tag{104}$$

Combining Eqn. (103) and Eqn. (104) via a union bound and using that $a+b \leq 2\max(a,b)$ completes the proof. □

This matches the bound proven in Portier et al. (2023, Proposition 3) which assumed that the test data was independent and identically distributed instead of on a regular grid. This rate of convergence is reasonably fast, particularly in low-dimensions. In particular ignoring the dependence of the bound on M^{test}, which is reasonable as the number of test points in map prediction is often far larger than the number of available points for training and validation, it is faster than the minimax

optimal rate of convergence for Lipschitz functions of $\theta((N^{\text{train}})^{-1/(d+2)})$ discussed earlier. We therefore would expect both spatial nearest neighbors and 1-nearest neighbors to perform well for model selection for grid prediction tasks if a fixed percentage of the data is used for training, and the remainder for validation. We emphasize we do not give a formal proof of this, just a heuristic argument suggesting why this should be the case. To give a formal proof would involve at least ensuring estimates of the risk estimation procedure hold uniformly over both sequences of predictive methods, and therefore involve additional assumptions.

D.6.4. General Prediction Tasks

For general Q^{test} we can combine Cor. D.14 together with Prop. 3.3 to get some sense of the rate of convergence of the spatial nearest neighbor method if the validation data is independent and identically distributed form a measure with density lower bounded on $[0,1]^d$ and the test task is fixed.

Corollary D.18 (Convergence of Spatial Nearest Neighbor with Independent and Identically Distributed Validation Data)**.** *Suppose that $\mathcal{S} = [0,1]^d$, $S_n^{\text{val}} \overset{\text{iid}}{\sim} P$ for $1 \leq n \leq N^{\text{val}}$ with $N^{\text{val}} > 1$ and P has Lebesgue density lower bounded by $c > 0$. Additionally, take the assumptions of Cor. D.14. Fix $\delta \in (0,1)$. Then there exists a constant $K_{d,\delta,L,\Delta,c}$ that depends only on d, δ, L, c and Δ such that with probability at least $1-\delta$*

$$|\hat{R}_{\text{NN},k}(h) - \hat{R}_{\text{NN},k^\star_{T_2}}(h)| \leq K_{d,\delta,L,\Delta,c} \left(\left(\frac{\log N^{\text{val}}}{N^{\text{val}}} \right) \right)^{\frac{1}{d+2}}. \tag{105}$$

Proof. From Prop. 3.3, there exists a constant γ depending on d, L, δ, c, Δ such that with probability at least $1 - \delta/2$

$$\tilde{\rho} \leq \gamma \left(\frac{\log N^{\text{val}}}{N^{\text{val}}} \right)^{1/d}. \tag{106}$$

Cor. D.14 implies that there exists a constant $K \geq 0$ such that with probability at least $1 - \delta/2$

$$|\hat{R}_{\text{NN},k}(h) - \hat{R}_{\text{NN},k^\star_{T_2}}(h)| \leq K \tilde{\rho}^{\frac{d}{d+2}}. \tag{107}$$

Combining Eqn. (106) and Eqn. (107) completes the proof. □

In this case, again up to logarithmic factors, Cor. D.18 means that SNN converges at the optimal rate of convergence for Lipschitz functions. This means we do not necessarily expect to be able to distinguish between two sequences of predictive methods that converge at the minimax rate, but we might expect to distinguish between two sequences of predictive methods if one converges to the optimal predictor at much slower than the minimax rate. We again emphasize that we do not formally show this, and to do so would would involve at least ensuring estimates of the risk estimation procedure hold uniformly over both sequences of predictive methods, and therefore involve additional assumptions.

In contrast, both the holdout and 1-nearest neighbor methods are not even always consistent for risk estimation in this setting, and therefore cannot be expected to reliably perform model selection.

D.6.5. Grid Prediction Proofs

In order to prove the claimed upper bound on grid prediction Cor. D.16 we use Thm. 5.1 together with an upper bound on the number of test points that lie within a ball of radius equal to the fill distance around any validation point. In order to do this, we will use that all the points in a grid are well-separated. We therefore begin by recalling the definition of a packing of a set, as well as a relationship between covering number and packing number.

Definition D.19 (Packing, Packing Number)**.** Let $A \subset \mathbb{R}^d$ a compact set. A (finite) set $B \subset A$ is called an ϵ-packing of A if for all $b, b' \in B$, $\|b - b'\| > \epsilon$. The ϵ-packing number of a set A, $\mathfrak{M}(\epsilon, A)$ is the largest cardinality of an ϵ-packing of A.

Proposition D.20 (Packing and Covering Numbers Wainwright 2019, Lemma 5.5)**.** *For any $A \subset \mathbb{R}^d$ and $\epsilon > 0$,*

$$\mathfrak{M}(2\epsilon, A) \leq \mathfrak{N}(\epsilon, A) \leq \mathfrak{M}(\epsilon, A). \tag{108}$$

We now restate and prove Cor. D.16.

Corollary D.16 (Bound on Estimation Error for Grid Prediction)**.** *With the same assumptions as Thm. 5.1 and additionally Assumption D.15, with probability at least* $1-\delta$

$$|R_{Q^{\text{test}}}(h) - \hat{R}_{\text{NN},1}(h)| \leq L\rho + \beta_\delta \sqrt{\max(\tfrac{2^d}{M^{\text{test}}}, (8\rho)^d)}. \tag{102}$$

Also, with probability at least $1-\delta$

$$|R_{Q^{\text{test}}}(h) - \hat{R}_{\text{NN},k^\star_{T_2}}(h)| \leq C_L\left(\rho + \beta_\delta \sqrt{\max(\tfrac{2^d}{M^{\text{test}}}, (8\rho)^d)}\right).$$

with $\rho = \rho_1, \beta_\delta = \Delta\sqrt{\frac{1}{2}\log\frac{2}{\delta}}$ *and* $C_L = \max(1, L)$.

Proof. The second inequality follows from the first by Prop. D.12 and because $1 \in T_2$. We therefore focus on proving the case $k = 1$.

In light of Thm. 5.1, it suffices to show that for these test location,

$$M^{\text{test}} Q^{\text{test}}(B(S_n^{\text{val}}, \rho)) = |B(S_n^{\text{val}}, \rho) \cap \{a/g : 1 \leq a \leq g\}^d| \leq \max\left(2^d, 8^d \rho^d M^{\text{test}}\right). \tag{109}$$

The set $\{a/g : 1 \leq a \leq g\}^d$ is a $\frac{1}{g+\epsilon}$-packing for any $\epsilon > 0$, and so by the first inequality in Prop. D.20 and Prop. D.3

$$|B(S_n^{\text{val}}, \rho) \cap \{a/g : 1 \leq a \leq g\}| \leq \mathfrak{M}(\tfrac{1}{(g+\epsilon)}, B(S_n^{\text{val}}, \rho)) \leq \mathfrak{N}(\tfrac{1}{2\rho(g+\epsilon)}, B(0,1)). \tag{110}$$

Applying Wainwright (2019, Lemma 5.7, Equation 5.9) and taking the limit as $\epsilon \to 0^+$

$$\lim_{\epsilon \to 0^+} \mathfrak{N}(\tfrac{1}{2(g+\epsilon)\rho}, B(0,1)) \leq \lim_{\epsilon \to 0^+} (1 + 4(g+\epsilon)\rho)^d = (1+4g\rho)^d. \tag{111}$$

By the binomial theorem and bounding the sum by the number of terms times the largest term

$$(1+4g\rho)^d \leq 2^d \max(1, 4^d g^d \rho^d) = \max(2^d, 8^d M^{\text{test}} \rho^d). \tag{112}$$

□

E. Additional Experimental Details

In this section, we provide additional details about the data, fitting procedures and validation procedures used in Section 6. Code used in experiments is available at: `https://github.com/DavidRBurt/Consistent-Spatial-Validation`. Code is almost all implemented in Python3 (Van Rossum & Drake, 2009) (with a small amount of r). Numpy is also heavily used for data generation and array manipulation (Harris et al., 2020).

In App. E.1 we give an overview of our method for estimating ground truth test risk in all experiments. In App. E.3 we describe details of the synthetic experiment described in Section 6.1. In App. E.4 we provide additional details on the air temperature data and tasks in Sections 6.2 and 6.5. In App. E.5 we provide additional details on the UK flat price prediction experiment described in Section 6.3, while in App. E.6 we provide additional details for the wind speed prediction task presented in Section 6.4.

E.1. Monte Carlo Estimation of Ground Truth Test Risk

We would like to compute the exact test risk across the M^{test} test points:

$$R_{Q^{\text{test}}}(h) := (1/M^{\text{test}}) \sum_{m=1}^{M^{\text{test}}} \mathbb{E}[\ell(Y_m^{\text{test}}, h^\chi(S_m^{\text{test}}))\big|S_m^{\text{test}}, \chi]. \tag{113}$$

In all our examples where we report test risk, we have access to some sample $(Y_m^{\text{test}})_{m=1}^{M^{\text{test}}}$ that we will use to construct an estimator. Our plan is to instead use the empirical test risk $\hat{R}_{Q^{\text{test}}}(h)$ as ground truth:

$$\hat{R}_{Q^{\text{test}}}(h) := (1/M^{\text{test}}) \sum_{m=1}^{M^{\text{test}}} \ell(Y_m^{\text{test}}, h^\chi(S_m^{\text{test}})). \tag{114}$$

We would like to know how far off the empirical test risk is from the exact test risk. To that end, we observe that

$$\hat{R}_{Q^{\text{test}}}(h) - R_{Q^{\text{test}}}(h) = (1/M^{\text{test}}) \sum_{m=1}^{M^{\text{test}}} Z_m, \quad \text{where} \tag{115}$$

$$Z_m := \ell(Y_m^{\text{test}}, h^\chi(S_m^{\text{test}})) - \mathbb{E}[\ell(Y_m^{\text{test}}, h^\chi(S_m^{\text{test}}))|S_m^{\text{test}}, \chi]. \tag{116}$$

By construction, if we assume the expectations exist, each random variable Z_m has mean zero. We make two additional assumptions. (1) We assume that the Z_m are independent. (2) We assume that (almost surely) $\forall m, Z_m \in (a, b)$ for finite $a, b \in \mathbb{R}$. If the loss is bounded by Δ, then such an a, b exist satisfying $b - a \leq \Delta$; following Assumption 2.1, we use this bound moving forward.

Under these assumptions, we can apply Hoeffding's inequality to conclude that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$|\hat{R}_{Q^{\text{test}}}(h) - R_{Q^{\text{test}}}(h)| \leq \Delta \sqrt{\frac{1}{2M^{\text{test}}} \log \frac{2}{\delta}}. \tag{117}$$

If we are willing to make the two assumptions above, we next show that we can reach (high probability) conclusions about the (true) relative quality of different estimators on a particular task if they pass a check: namely, we check if, for a small δ (e.g. $\delta = 0.05$), the right-hand side of Eqn. (117) is smaller than twice the difference between how much closer the "good" estimator is to the estimate of ground truth than the "bad" estimate. To see why this check is sufficient, first observe the following two applications of the triangle inequality:

$$|\text{good} - \text{true}| \leq |\text{good} - \widehat{\text{true}}| + |\text{true} - \widehat{\text{true}}| \tag{118}$$

$$|\text{bad} - \widehat{\text{true}}| \leq |\text{bad} - \text{true}| + |\text{true} - \widehat{\text{true}}|. \tag{119}$$

Using these two inequalities, we can write

$$|\text{bad} - \text{true}| - |\text{good} - \text{true}| \geq |\text{bad} - \widehat{\text{true}}| - |\text{good} - \widehat{\text{true}}| - 2|\text{true} - \widehat{\text{true}}|. \tag{120}$$

Therefore, to conclude

$$|\text{bad} - \text{true}| - |\text{good} - \text{true}| \geq 0, \tag{121}$$

it suffices for

$$|\text{bad} - \widehat{\text{true}}| - |\text{good} - \widehat{\text{true}}| \geq 2|\text{true} - \widehat{\text{true}}|. \tag{122}$$

Under the earlier assumptions, we see that Eqn. (122) is implied (with high probability) by

$$|\text{bad} - \widehat{\text{true}}| - |\text{good} - \widehat{\text{true}}| \geq 2\Delta \sqrt{\frac{1}{2M^{\text{test}}} \log \frac{2}{\delta}}. \tag{123}$$

When discussing each experiment, we discuss the plausibility of the assumptions needed to make this argument when justifying our estimated ground truth, as well as specific values for Δ and M^{test} and the resulting bound.

E.2. Computational Considerations

E.2.1. Computational Complexity of our Method

We focus on the case $\mathcal{S} = [0, 1]^d$ with nearest-neighbors implented using a kd-tree. Computation of the approximate fill distance is already discussed in App. D.5.4 and is shown to be $O(dN^{\text{val}} \log N^{\text{val}})$. Construction of a kd tree on the validation data is also $O(dN^{\text{val}} \log N^{\text{val}})$. Once constructed, finding the nearest neighbor for each test point requires $O(\log N^{\text{val}})$ computations of a d-dimensional Euclidean distance, meaning finding the neighbors is $O(dM^{\text{test}} k \log N^{\text{val}})$. Since $O(\log N^{\text{val}})$ values of k are tried in selecting k, this leads to a complexity not more than $O(dM^{\text{test}} N^{\text{val}} (\log N^{\text{val}})^2)$. We expect further improvements could be made by storing the nearest neighbor set for intermediate values of k, but we do not pursue these.

E.2.2. COMPUTATIONAL SETUP USED

All experiments were run on a CPU cluster with 36 Intel(R) Xeon(R) W-2295 CPU @ 3.00GHz CPUs and a total of 251 GB of system RAM. In all experiments linear algebra operations operations were allowed to be multithreaded, and so at times all 36 CPUs were used, even if fewer than 36 parallel jobs were run.

E.2.3. COMPUTATIONAL COST OF SYNTHETIC EXPERIMENT

Data Generation. Generating the point prediction synthetic data takes around 25 minutes using 10 parallel jobs and has a peak memory usage of around 46GB.

Generating the grid prediction synthetic data takes around 33 minutes using 10 parallel jobs and has a peak memory usage of around 71GB.

Running Experiment. Running the point prediction task takes around 25 minutes using 10 parallel jobs and has a peak memory usage of around 32GB.

Running the grid prediction task takes around 27 minutes using 10 parallel jobs and has a peak memory usage of around 39GB.

E.2.4. COMPUTATIONAL COST OF BOOTSTRAPPED AIR TEMPERATURE EXPERIMENT

Data Generation Running the make file to download air station data takes on the order of 30 seconds and not more than 6GB of RAM after some data has been installed manually as described in the README file in the released code. Fitting the model and computing residuals in order to generate bootstrapped datasets takes around 10.5 minutes and has peak memory usage around 95GB.

Running Experiment Running the bootstrapped air temperature metro prediction task takes around 66 minutes with 20 parallel jobs and has peak memory usage around 225GB.

Running the bootstrapped air temperature grid prediction task takes around 8.5 hours with 3 parallel jobs and has peak memory usage around 150GB.

E.2.5. COMPUTATIONAL COST OF UK HOUSE PRICE EXPERIMENT

Data Processing Downloading and processing the data takes around 7 seconds and has a peak memory usage of under 3GB.

Running Experiment Running the UK House price prediction task takes around 4.6 hours with 5 parallel jobs and has peak memory usage around 70GB.

E.2.6. COMPUTATIONAL COST OF WIND SPEED

Data Processing Downloading and processing the data takes around 3.5 minutes and has a peak memory usage of under 12GB.

Running Experiment Running the wind spped prediction task takes around 8.5 hours with 15 parallel jobs and has peak memory usage around 15GB.

E.2.7. COMPUTATIONAL COST OF REAL DATA AIR TEMPERATURE EXPERIMENT

Time to process the data has been previously described in App. E.2.4.

Running Experiment Running all the air temperature tasks takes around 10 minutes and has peak memory usage around 9GB.

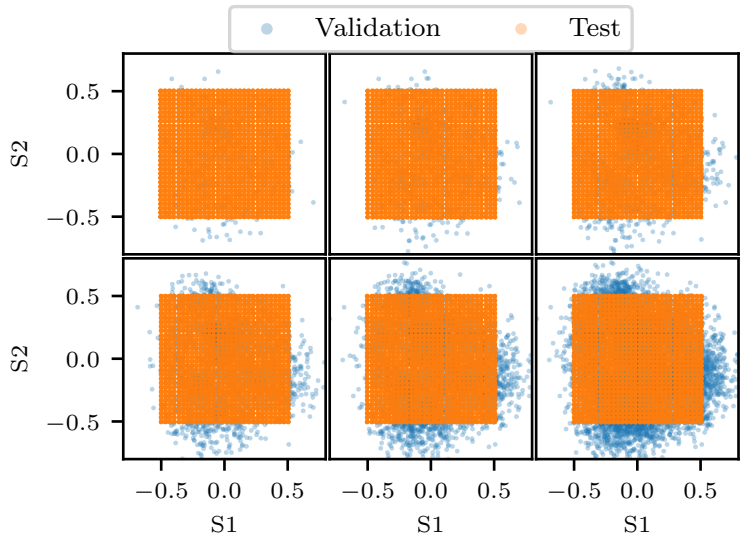

Figure 3: The validation sites (blue, clustered) for the first seed of the synthetic grid task. Panels from left to right and top to bottom represent N^{val} in the sequence $(250, 500, 1000, 2000, 4000, 8000)$. Test sites (orange, gridded) are constant across panels.

E.2.8. Computational Cost of Model Selection Experiment

Generating data for and running the synthetic model selection experiment takes under a minute of time and under 4GB of RAM.

E.3. Risk Estimation on Synthetic Data

In this section, we provide additional details and figures for our risk estimation experiment on synthetic data presented in Section 6.1. App. E.3.1 describes the process by which we generate both datasets considered and shows an example of the covariate and response spatial fields for each problem (Figs. 4 and 6). App. E.3.3 describes the procedure used to fit the predictive methods to each dataset. App. E.3.4 describes the implementation of the risk estimation procedures we compare. App. E.3.6 describes the metric reported, and Figs. 5 and 7 show the (signed) relative error of each estimator as we vary the amount of validation data available.

E.3.1. Simulation Data Generating Process Details

For both tasks, 100 datasets are generated following the process outlined below.

Grid data *Generation of training and validation sites:* The first training point is selected via generating a point uniformly in $[-0.5, 0.5]^2$, and making this the mean of a Gaussian mixture component, with standard deviation randomly sampled between 0.05 and 0.15. This mixture is initially given weight 1, the first training point is then sampled from a Gaussian with this mean and standard deviation, the and weight of this mixture is increased to 2. Subsequent points are sampled sequentially. For each i between 2 and the total number of training and validation points, a weight of 1 is assigned to adding a mixture component. The new point is then sampled from either one of the existing mixture components, or the new mixture component, with probability proportional to the current weights assigned to each mixture component. The weight of the mixture from which the points, $w_{i(t)}^{(t)}$, is then increased as

$$w_{i(t)}^{(t+1)} = w_{i(t)}^{(t)} + \frac{1}{w_{i(t)}^{(t)}}. \tag{124}$$

This is reminiscent of a Chinese restaurant process (Pitman, 2006, Section 3.1), but the weights are increased more slowly, leading to more clusters being formed and less large clusters typically.

If a new mixture component is generated, a mean for the mixture component is generated on $[-0.5, 0.5]^2$, and a standard deviation is selected uniformly on $[0.05, 0.15]$. Conditional on the mixture component, the new point is sampled from a Gaussian distribution with the components mean and standard deviation.

The first 1000 points generated this way are taken to be the training data, and the remaining N^{val} points generated this way are the validation data. An example of the training and validation data generated through this process are shown in the top left of Fig. 4.

Generation of test data The test data is $\{(-0.5 + a/29, -0.5 + b/29) : 0 \leq a, b \leq 49\}$. That is, it is a regular grid on $[-0.5, 0.5]^2$. We generate 50 values of each response variable on each. grid point.

Generation of Covariates

The covariates are generated as a zero-mean Gaussian process with an isotropic Matérn 3/2 covariance function with lengthscale 0.3 and scale parameter 1. That is, the covariance function is,

$$k_\chi(S, S') = \left(1 + \frac{\sqrt{3}\|S - S'\|_2}{0.3}\right) \exp\left(-\frac{\sqrt{3}\|S - S'\|_2}{0.3}\right). \tag{125}$$

A small diagonal term (1e-12) is added to the diagonal of the covariance matrix to avoid numerical linear algebra errors. Sampling is performed using Tensorflow probability (Dillon et al., 2017). We generate two covariates spatial processes via this process $\chi = (\chi^{(1)}, \chi^{(2)})$.

Generation of Response Once the sites and covariates have been generated, the response variable is sampled from a Gaussian process with zero mean. The covariance function of the Gaussian is a sum of two, 2 dimensional isotropic Matérn 3/2 kernels:

$$k(S, S') = 0.5\left(1 + \frac{\sqrt{3}\|S - S'\|_2}{0.5}\right) \exp\left(-\frac{\sqrt{3}\|S - S'\|_2}{0.5}\right) \tag{126}$$

$$+ \left(1 + \sqrt{3}\|\chi(S) - \chi(S')\|_2\right) \exp\left(-\sqrt{3}\|\chi(S) - \chi(S')\|_2\right). \tag{127}$$

Independent, identically distributed Gaussian noise is added to the function values with variance 0.1.

Point Prediction Task *Generation of training and validation sites:* The training and validation points are sampled independently and identically from a uniform distribution supported on $[-0.5, 0.5]^2$. 1000 training points are used in all experiments. The number of validation points is varied in $\{250 \times 2^\ell\}_{\ell=0}^5$.

Generation of test site The test site is fixed to be the origin. We generate 45000 response values at the origin so that when we compute the empirical risk we expect it to accurately reflect that actual risk.

Generation of covariates: The covariates are generated as a zero-mean Gaussian process with an isotropic squared exponential covariance function with lengthscale 0.3 and scale parameter 1. That is, the covariance function is,

$$k_\chi(S, S') = \exp\left(-\frac{\|S - S'\|_2^2}{2 \cdot 0.3^2}\right). \tag{128}$$

A small diagonal term (1e-12) is added to the diagonal of the covariance matrix to avoid numerical linear algebra errors. Sampling is performed using tensorflow probability (Dillon et al., 2017). We generate two covariates via this process $X = (X^{(1)}, X^{(2)})$.

Generation of response Once the sites and covariates have been generated, the response variable is sampled from a Gaussian process with zero mean. The covariance function of the Gaussian is a sum of two, 2 dimensional isotropic squared exponential kernels:

$$k(S, S') = 0.5 \exp\left(-\frac{\|S - S'\|_2^2}{2 \cdot 0.5^2}\right) + \exp\left(-\frac{\|\chi(S) - \chi(S')\|_2}{2}\right). \tag{129}$$

Independent, identically distributed Gaussian noise is added to the function values with variance 0.1.

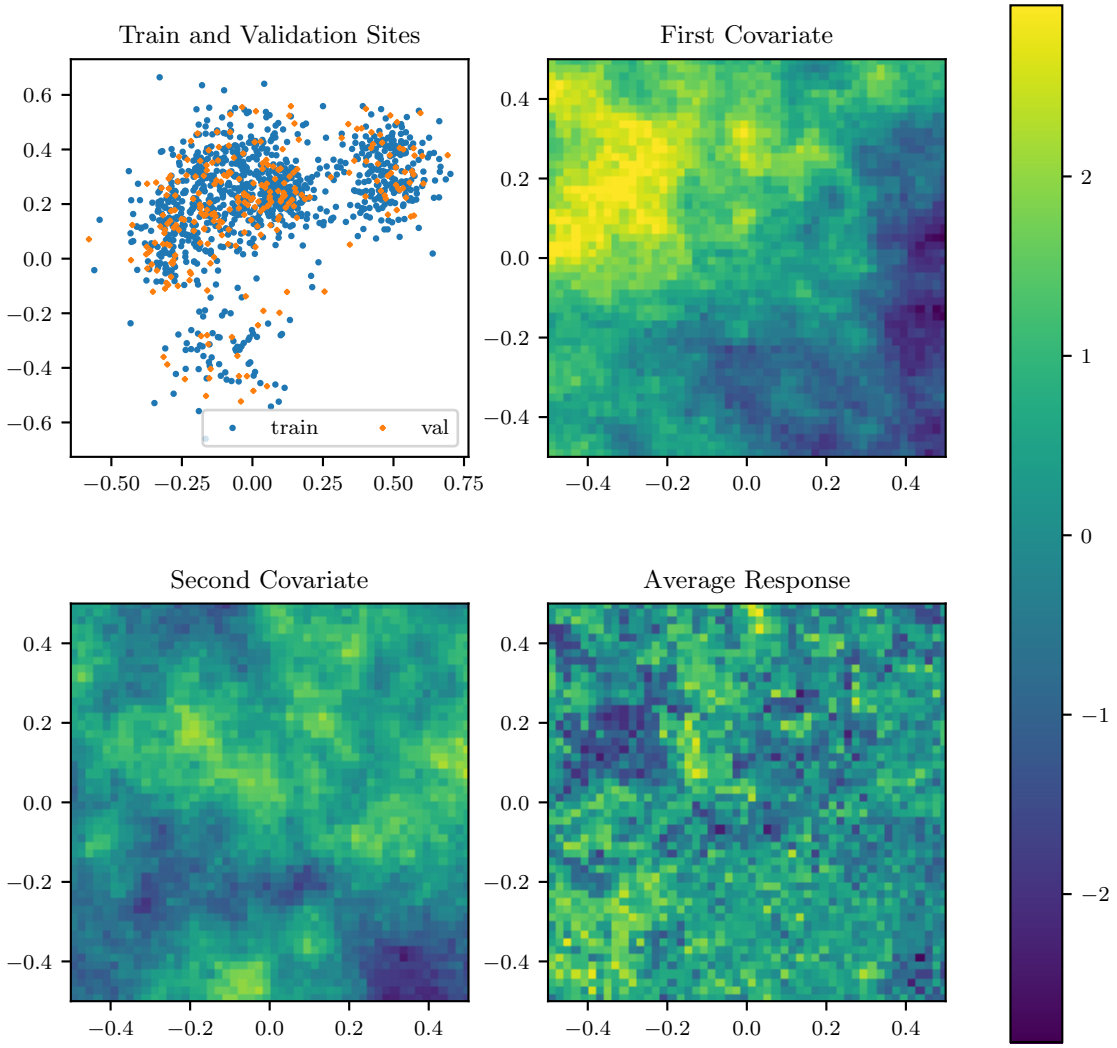

Figure 4: Data for the grid prediction task. An example of a single sample of the training and validation points (with 1000 training points and 250 validation points) is shown in the top left. The top right and bottom left show the covariates as a function of space, while the bottom right shows the mean of the response variable as a function of space.

Table 2: Value of $k^\star_{T_2}$ chosen by minimizing the bound in the grid prediction task. Because the test data is well separated, the variance of the estimator is small even when k is small. As a result the value of k that minimizes the upper bound is generally small, even as the number of validation points increases.

	Number of Validation Points					
$\mathbf{k^\star_{T_2}}$	250	500	1000	2000	4000	8000
1	82	85	81	82	66	50
2	15	15	19	18	34	41
4	3	0	0	0	0	9

Table 3: Value of $k^\star_{T_2}$ chosen by minimizing the bound in the point prediction task. In this task, there is generally a bias-variance trade-off that must be balanced. As a result the value of k that minimizes the upper bound increases as the amount of available validation data increases.

	Number of Validation Points					
$\mathbf{k^\star_{T_2}}$	250	500	1000	2000	4000	8000
16	7	0	0	0	0	0
32	93	33	0	0	0	0
64	0	67	99	14	0	0
128	0	0	1	86	98	0
256	0	0	0	0	2	100

E.3.2. Loss Function

We use truncated, squared loss,

$$\ell(a, b) = \min(1.0, (a - b)^2), \tag{130}$$

which is bounded by 1.0. The empirical risk is calculated as in Eqn. (114).

E.3.3. Model Fitting

We fit a Gaussian process regression model to using only the first covariate $\chi^{(1)}$ to the training data. The prior is taken to be the same as the data generating process, but with (only) $\chi^{(1)}$ in place of $(\chi^{(1)}, \chi^{(2)})$ in the second kernel in Eqn. (127) and Eqn. (129) for the two datasets respectively. The mean of the posterior process is used for predictions, and is calculated using GPFlow (Matthews et al., 2017).

E.3.4. Implementation of Risk Estimation

The holdout is implemented by taking an (unweighted) average of the loss on each validation point. Both nearest neighbor methods are implemented using `scikit-learn` (Pedregosa et al., 2011) with kd-trees and Euclidean distance. For $k^\star_{T_2}$ nearest neighbors, nearest neighbors is performed for all k that are powers of 2 less than N^{val}, and the value of k with the smallest bound is used for risk estimation. This is done with $\delta = r$, with r calculated as in App. D.5.4, and $\Delta = 1$. Table 2 and Table 3 show the values of k chosen for grid and point prediction respectively. For the grid prediction task, $k^\star_{T_2}$ tends to be small, even as the size of the validation set becomes large. This supported by Cor. D.16, since even 1-nearest neighbor reliably estimates risk in this setting. For the point prediction task, $k^\star_{T_2}$ grows with N^{val}.

E.3.5. Monte Carlo Estimation of Test Risk

Following the argument in App. E.1, we use the empirical test risk Eqn. (114) in place of the test risk as ground truth in synthetic experiments. The assumption that $(Z_m)_{m=1}^{M^{\text{test}}}$ are independent holds by the description of the data generating process, because the $(Y_m)_{m=1}^{M^{\text{test}}}$ are conditionally independent and Z_m is a function of Y_m. The assumption that Z_m is almost surely bounded holds with $\Delta = 1$ by our choice of truncated squared loss. Further, in both synthetic experiments, we take

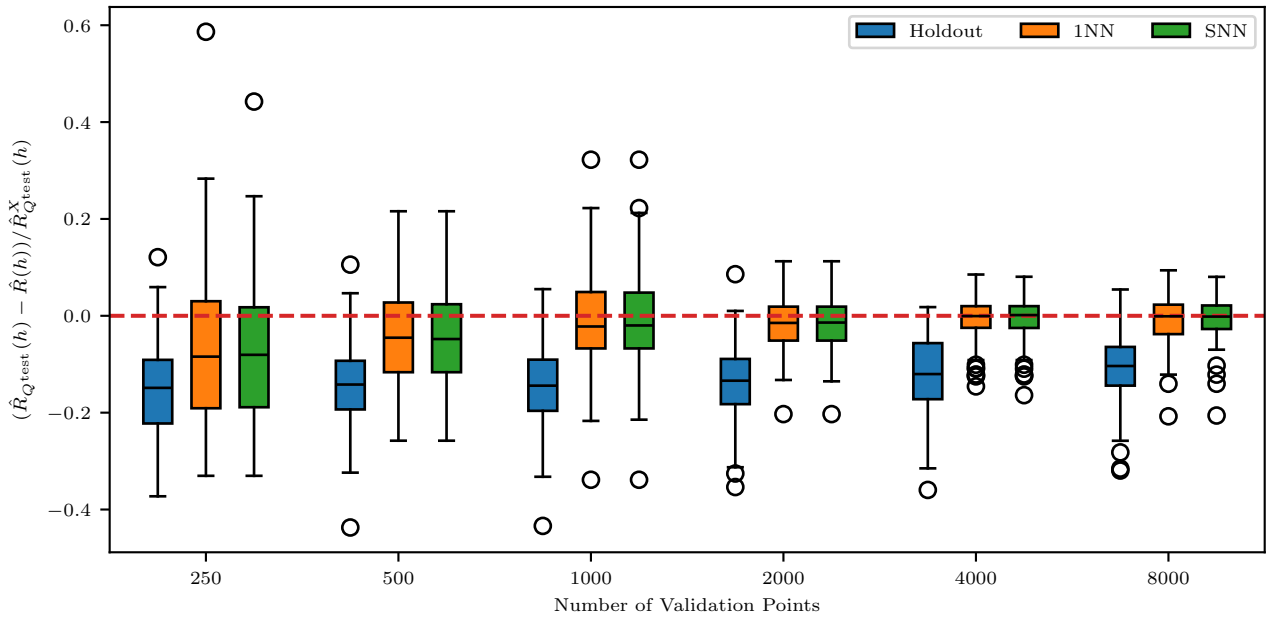

Figure 5: The relative error in estimating the empirical risk for each method is plotted against the number of validation points used for the grid prediction task. The holdout is biased, even for large N^{val}. The 1-nearest neighbor and $k^\star_{T_2}$-nearest neighbor estimates both have small relative error for large N^{val}.

$M^{\text{test}} = 45000$. Combining these gives that with probability at least 0.95

$$|\hat{R}_{Q^{\text{test}}}(h) - R_{Q^{\text{test}}}(h)| \leq \sqrt{\frac{1}{2 \times 45000} \log \frac{2}{0.05}} \leq 0.0065. \tag{131}$$

Fig. 1 shows the absolute difference between each estimator and the empirical test risk across 100 seeds. We see that the difference between the estimators is generally larger than twice Eqn. (131), and so by the argument in App. E.1, we expect our estimate of ground truth to be accurate enough that the difference in performance of the methods is not simply due to error in estimating the ground truth.

E.3.6. Metrics Reported and Additional Figures

Figs. 5 and 7 show the relative errors of each estimation, calculated as

$$\frac{\hat{R}_{Q^{\text{test}}}(h) - \hat{R}(h)}{\hat{R}_{Q^{\text{test}}}(h)}. \tag{132}$$

From this, we can see that the holdout method has a bias in both cases that does not appear to go away as the number of validation points increases. In contrast, the 1-nearest neighbor method primarily suffers due to a variance issue when it fails to converge. We again see in both instances the $k^\star_{T_2}$-nearest neighbor approach appears to concentrate around zero error as the number of validation points increases.

E.4. Air Temperature Tasks

We now provide additional details about data source, pre-processing, model fitting and risk estimation for the air temperature dataset. These are identical between the real response experiment and the partially synthetic experiment, except for the bootstrapping procedure described in App. E.4.6. We also provide additional experimental results on a grid prediction task for both the bootstrapped and original datasets.

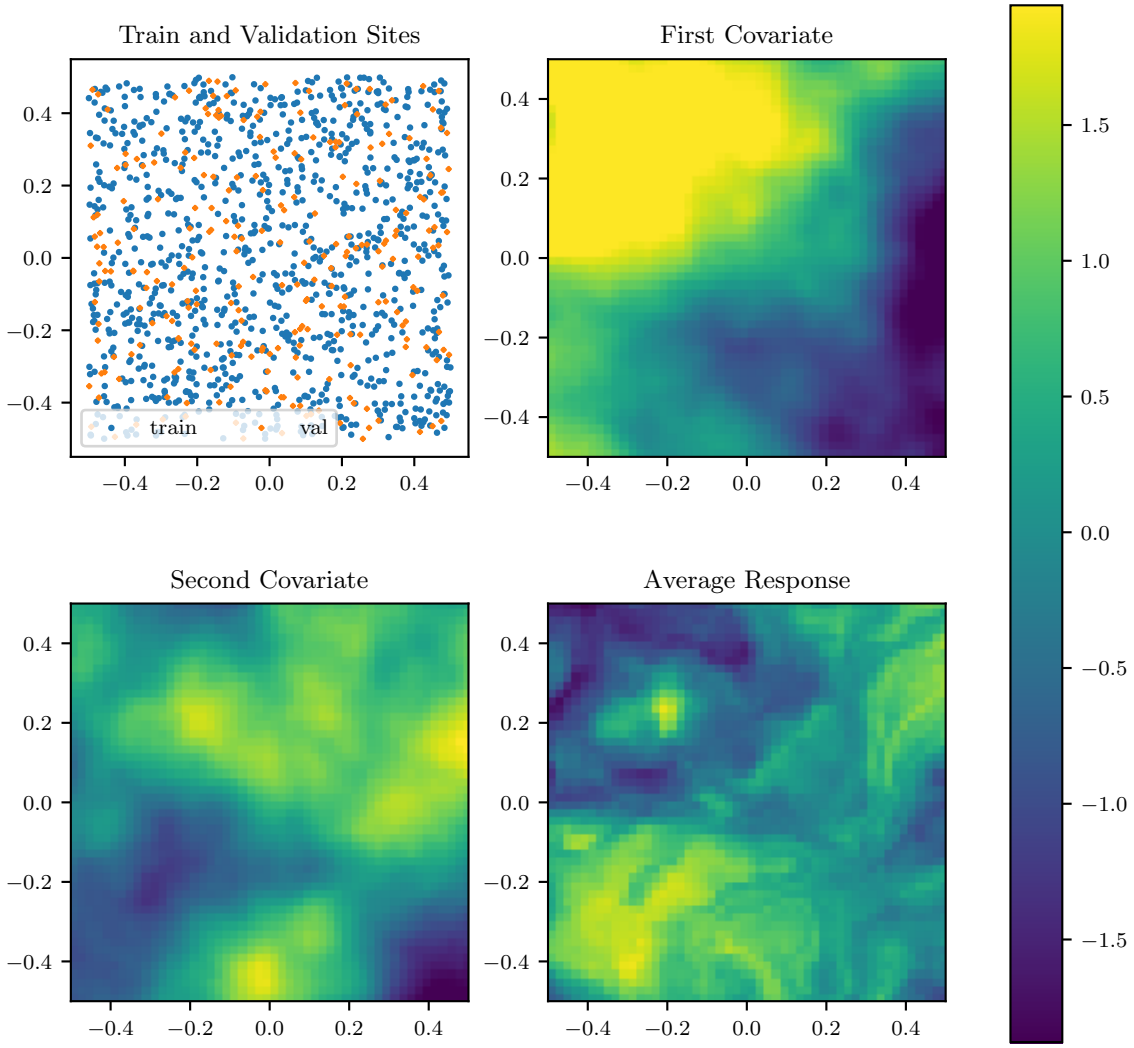

Figure 6: Data for the point prediction task. An example of a single sample of the training and validation points (with 1000 training points and 250 validation points) is shown in the top left. The top right and bottom left show the covariates as a function of space, while the bottom right shows the mean of the response variable as a function of space.

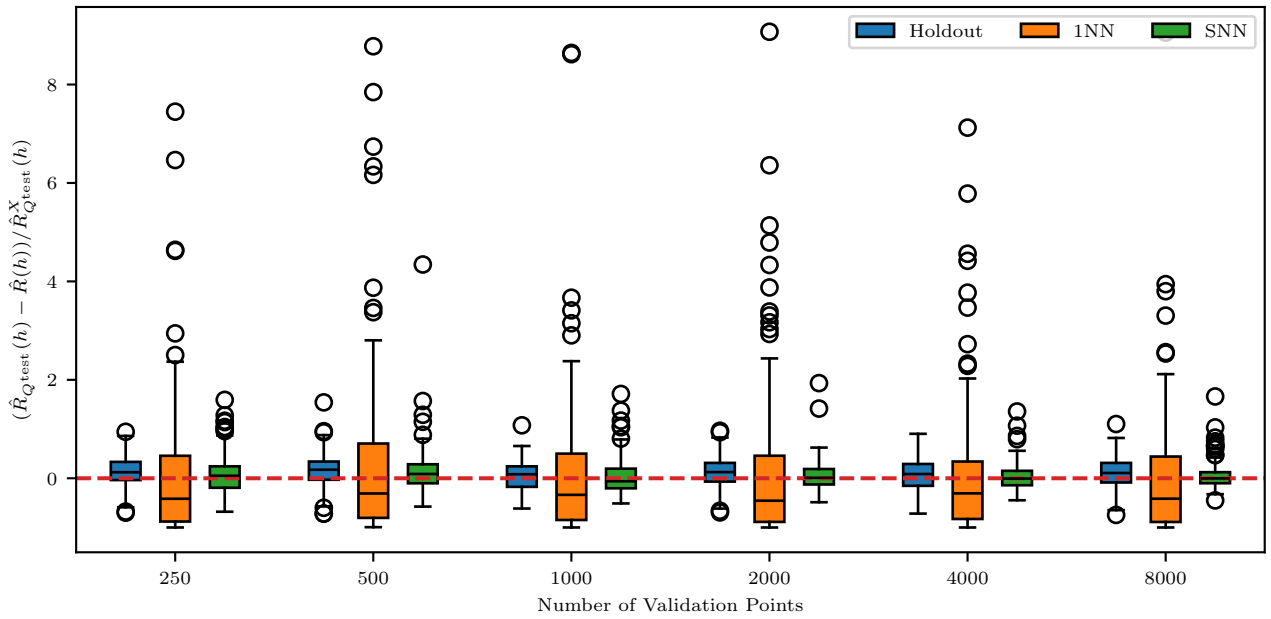

Figure 7: The relative error in estimating the empirical risk for each method is plotted against the number of validation points used for the point prediction task. The holdout has a small but non-negligible bias. The 1-nearest neighbor method has a small bias, but large variance. The nearest neighbor procedure with $k^\star_{T_2}$ neighbors has small relative error for large N^{val}.

E.4.1. Data Sources

The land surface temperature used is from MODIS Aqua (Wan et al., 2021) and is monthly average land surface temperature on a 0.05 degree grid. We download monthly average weather station data from the Global Historical Climatology Network (Menne et al., 2018). Latitude and longitude of major United States (US) urban areas are from the 2023 US census gazetteer (United States Census Bureau, 2023). All of these datasets are produced in large part by US government agencies (NASA, NOAA and the Census). While we could not find specific license information, we understand these datasets to be public domain following section 105 of the Copyright Act of 1976.

E.4.2. Data Pre-processing

We assign the land surface temperature at the nearest point (using a spherical approximation to distance between points) to each weather station. The nearest point is found using the `scikit-learn` implementation of nearest neighbor algorithm using the 'ball-tree' (Omohundro, 2009) and 'Haversine' metric. Temperatures are converted to Celsius from Kelvin. We remove all rows where the Land Surface temperature is not available. We use the weather station data uploaded to GHCNM as of January 15, 2024. We filter out weather stations outside of the United States (based on the station ID). We also filter out stations with a non-empty quality control flag or no temperature recorded for January 2018 (the month we consider). Finally, we remove stations in Hawaii or Alaska to focus on the continental United States. In total, after this processing, there are 6422 weather stations. We use 70% of the stations for fitting the models, and holdout the remaining 30% estimates. When building the test sites, we remove points outside the United States based on a reverse geocoding lookup with Thampi (2015) to the nearest city. This does not create an exact boundary (since it is based on the nearest city or town and not the country in which the latitude and longitude is based in) but is a good proxy for whether or not a point is in the United States. Fig. 8 of the available weather stations for model fitting and validation, colored by monthly average temperature in January 2018.

E.4.3. Loss Function

We consider a truncated absolute value as the loss function, $\ell(a, b) = \min(5.0, |a - b|)$. This means we are primarily interested in the quality of the model predictions when it is relatively close to the actual response, and do not consider

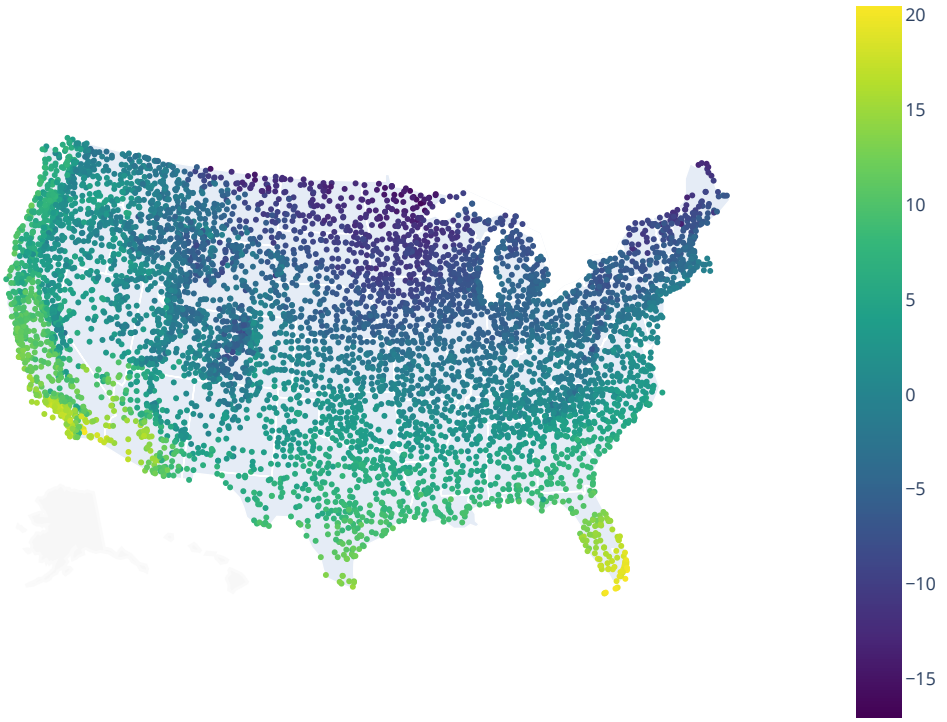

Figure 8: Weather stations used in the air temperature prediction task we considered, colored by average temperature in January 2018 in degrees Celsius.

differences in predictions that are, for example, 8 degrees Celsius versus 10 degrees Celsius wrong meaningfully different. While the choice of 5 degrees is arbitrary, this is motivated by applications in which we might have some allowable tolerance for the quality of a prediction beyond which the prediction is no longer useful (and so it doesn't matter how bad it is).

E.4.4. Model Fitting

Inspired by Hooker et al. (2018), we fit a geographically weighted least squares regression model using the land surface temperature at day and night. In particular, we fit an affine model, with the coefficients, $\beta(S)$ depending on the location that will be predicted at. $\beta(S)$ is selected by solving the weighted least squares problem,

$$\hat{\beta}(S) \in \arg\min_{(b_0,b_1,b_2)\in\mathbb{R}^3} \sum_{i=1}^{n^{\text{train}}} w_i(S)(Y_i^{\text{train}} - (b_0 + b_1 X_i^{\text{train},1} + b_2 X_i^{\text{train},2}))^2, \tag{133}$$

with Y_i^{train} the temperature at station i, $X_i^{\text{train},1}$ the daytime land surface temperature $X_i^{\text{train},2}$ the nighttime land surface temperature both at the closest satellite point to station i and $w_i(S) = \exp(-\frac{d_{\text{haversine}}(S,S_i^{\text{train}})^2}{2\ell^2})$ and $d_{\text{haversine}}$ the Haversine (great circle) distance between the points. $\ell \geq 0$ is a parameter, and we select it from $\{25.0, 50.0, 75.0, 100.0, 150.0, 200.0, 300.0, 400.0, 500.0, 750.0, 1000.0\}\text{km}$ via leave-one-out cross-validation on the training data with mean squared error. We perform leave-one-out cross-validation (without additional weighting).

We also consider a simple baseline model fit using only the weather station data. We fit a Gaussian process with zero prior mean and Matérn 3/2 kernel to the weather stations with covariate the spatial locations in latitude and longitude converted to radians, and a Gaussian likelihood model. We fit the parameters of the kernel using L-BFGS to attempt to maximize the marginal likelihood of the parameters. The parameters fit are two lengthscale parameters (one for each spatial dimension), a kernel scale parameter, and a likelihood variance parameter. The mean is removed from the training data prior to fitting, the kernel lengthscales are set to to standard deviation of each covariate and the kernel variance parameter is set to equal the variance of the training response data, and the likelihood variance parameter is set to equal 0.1-times the variance of the training response variable. A maximum of 15 iterations of L-BFGS are run.

E.4.5. Risk Estimation Details

The holdout is implemented as in previous experiments (App. E.3.4). We estimate the standard error of the holdout empirically by computing the sample standard deviation of the sum of the losses,

$$\hat{\sigma}^2 = \left(\frac{1}{N^{\text{val}}(N^{\text{val}} - 1)} \sum_{j=1}^{N^{\text{val}}} \ell(Y_j^{\text{val}}, h^{\chi}(S_j^{\text{val}}))^2 \right)^{1/2}. \tag{134}$$

Table 1 reports the holdout estimate $\pm$ two standard deviation.

The nearest neighbor methods are implemented using the `scikit-learn` implementation with Haversine distance and the ball-tree algorithm. $k^{\star}_{T_2}$ is selected with $\delta = 0.1$ and $\Delta = 5°$C and a Lipschitz constant of $1°$C/100 km. We use a fixed δ since we only discuss a method applicable to estimating fill distance on the unit cube, not on a subset of the sphere.

E.4.6. Bootstrapping of Residuals

In order to generate many datasets with a realistic synthetic response variable where we have access to ground truth we:

1. Fit a Gaussian process regression model to *all* the available weather station data. We use a Matérn 3/2 kernel with zero prior mean on the weather station data with the (spatial) mean temperature removed. Parameters of the kernel are selected via maximum likelihood.
2. Compute the empirical distribution of the residuals of the mean of these predictions.
3. For each seed we then use the same spatial locations and covariates, and generate the response surface at any point in space by computing the mean of the Gaussian process regression model fit and adding a sample from the empirical distribution of the residuals of the actual data.

We can then directly estimate the test risk via generating many Y^{test} at each spatial location (we use 1000 realizations for each city in the 5-metros task) and 1 for each grid point in the grid task (since the error is averaged over test sites this still results in an estimator that is concentrated) in this manner and forming a Monte Carlo estimate as in App. E.1.

E.4.7. Estimation of Ground Truth in Bootstrapped Experiment

Following the argument in App. E.1, we use the empirical test risk Eqn. (114) in place of the test risk as ground truth in synthetic experiments. For the assumption that $(Z_m)_{m=1}^{M^{\text{test}}}$ are independent to hold it is sufficient for Y_m^{test} to be independent, conditioned on the spatial location at which it is observed. This holds based on the data generating process used to construct the synthetic responses: since Y_m^{test} is a noisy observation of the smooth function we fit to the weather stations, plus noise sampled independently from the distribution of residuals. Because the loss is truncated MAE, the Z_m are surely bounded by 5. We use 10000 samples at each of the 5 test location in estimating the risk. Using these numbers in Eqn. (117), we arrive at

$$|\hat{R}_{Q^{\text{test}}}(h) - R_{Q^{\text{test}}}(h)| \leq 5\sqrt{\frac{1}{2 \times 50000} \log \frac{2}{0.05}} < 0.031. \tag{135}$$

Following the argument in App. E.1, we expect our estimate of ground truth to be good enough to distinguish between the quality of models whose absolute error from the estimated ground truth differs by more than $2 \times 0.031 = 0.062$. Fig. 9 shows these absolute errors. We see that for many of the seeds, the difference between the performance of 1NN (orange) and SNN and the holdout is greater than 0.174. Given that the earlier argument is quite conservative (in the sense that Hoeffding's inequality is likely to be loose), we therefore can attribute the difference in performance of the methods to indicate that SNN and the holdout are giving better estimates of the ground truth test risk, and the observed difference is not due to error in our estimation of the test risk.

For the grid prediction task the assumptions are similar, but $M^{\text{test}} = 341{,}628$ as this is the number of points on the map. Because the assumptions are satisfied by construction of the synthetic data (as in the 5-metro prediction task), with probability $1 - \delta$

$$|\hat{R}_{Q^{\text{test}}}(h) - R_{Q^{\text{test}}}(h)| \leq 5\sqrt{\frac{1}{2 \times 341628} \log \frac{2}{0.05}} < 0.012. \tag{136}$$

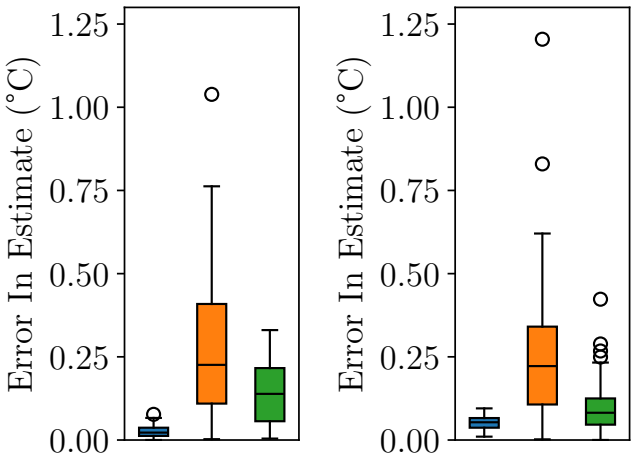

Figure 9: Absolute error in estimating the truncated mean absolute error in the air temperature dataset with bootstrapping.

Table 4: Estimates of risk given by each method. All three methods agree reasonably well (within ± 2 standard deviation of the estimate given by holdout for both geographically weighted regression and spatial regression in this task. In particular, all three methods suggest that the geographically weighted regression method has lower risk on this task.

	GWR	Spatial GP
Holdout	$\mathbf{0.83 \pm 0.03}$	0.90 ± 0.04
1NN	**0.80**	0.88
SNN	**0.80**	0.88

We therefore expect our estimate of ground truth to be very accurate, although we see in Fig. 10 that the methods all perform well in estimating the test risk on this task, and so there is likely not a meaningful difference in which approach is used to perform validation.

E.4.8. Results for Grid Prediction with Bootstrapped Data

Fig. 10 shows the results for holdout, 1-nearest neighbor and SNN with the test set each grid point in the map that is located in the continental United States. All 3 methods lead to reasonably accurate estimates of the mean absolute error on this prediction task (within 0.1 degrees of the ground truth error). Based on our theory, we generally expect 1NN and SNN to have small error in grid prediction tasks (at least with sufficient data and the infill assumption being satisfied), while for the holdout it depends on the particular predictive method and distribution of test and validation sites. In this case, it appears for both prediction methods the bias introduced by the use of the holdout is relatively small.

E.4.9. Results for Grid Prediction with Real Data

Table 4 shows the results for holdout, 1-nearest neighbor and SNN with the test set each grid point in the map that is located in the continental United States. We see good agreement between all three method. This is expected for 1-nearest neighbor and SNN based on earlier theory (App. D.6.5).

E.5. UK Housing Experiment

We provide additional details for the UK flat price prediction task presented in Section 6.3.

E.5.1. Data Sources and Pre-processing

We download 2023 price paid data for England and Wales from HM Land Registry (2023). This data is subject to a UK Open Government License (https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/), which requires citation of the data, but allows both commercial and

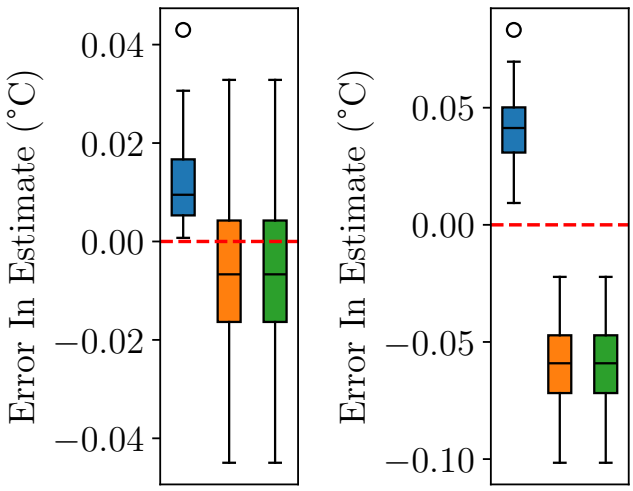

Figure 10: Error in estimate of mean absolute error of the Gaussian process regression predictions (left) and the geographically weighted regression predictions (right) for the holdout (blue), 1NN (oragne) and SNN (green) on a grid prediction task. We see that all 3 methods result in accurate estimates (within 0.1 degrees Celsius) of the mean absolute error on this task for both prediction methods, and suspect it is unlikely meaningfully different conclusions would be drawn from use of any of the methods in this application.

non-commercial uses. These records contain postal codes for each property sold, the type of property sold, town or city, price paid for the property. We use the type of property variable to filter out all properties that are not flats, and only consider additions (not replacements or deletions) to the dataset and "standard" price paid data (not repossessions, buy-to-lets or other sales labelled as non-standard). Noting that a postal code in the UK corresponds to a very small geographic area, we obtain latitude and longitude data for each sale by looking up the postal code coordinates using the National Statistics Postcode Lookup (Office Of National Statistics, 2024), which we understand to be a product of the UK Census and therefore also subject to an Open Government License. We convert from northing and easting to latitude and longitude using `R`. We log transform the price variable prior to model fitting as we expect price paid to be non-negative and highly skewed, and so a Gaussian (process) prior would otherwise be almost certainly inappropriate.

E.5.2. Model Fitting

We fit hyperparameters of the variational Gaussian process regression by evidence lower bound maximization.

Model Specification We fit a Gaussian process regression model with prior covariance specified by a sum of two Matérn 3/2 kernels and a zero prior mean on the mean centered log price paid data. We use a sum of Matérn 3/2 kernel in place of the sum of RBF kernel used in Hensman et al. (2013) as we expect there to be places where (log) property prices vary quickly in space, and so the smoothness properties implicitly assumed in using an RBF kernel may be inappropriate. We use 2000 inducing points for the variational approximation. The locations of these points are optimized jointly with model parameters when maximizing the evidence lower bound. We use the closed form for the optimal variational posterior (given a set of inducing points) provided in Titsias (2009), and perform maximization of the evidence lower bound using L-BFGS.

Initialization The locations of the inducing points are initialized by the greedy procedure suggested in Burt et al. (2020), which is essentially equivalent to a partially pivoted Cholesky decomposition recommended earlier in the Gaussian process approximation literature (Foster et al., 2009). The initial prior variance of both kernels is set to be equal to the variance in the training data; the lengthscales of one kernel (intended to model regional price trends) are initialized to twice the standard deviation in the location data (in radians), while the scale of the other kernel (intended to model local trends) is initialized to half the standard deviation in the location data. The likelihood standard deviation is initialized to be $0.1\times$ the standard deviation in the log price paid in the training data.

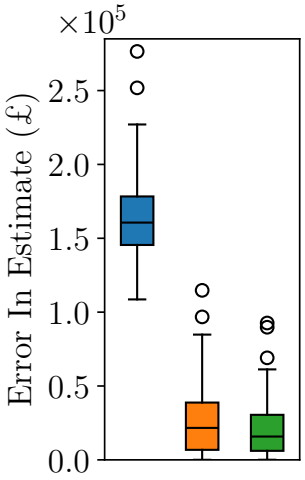

Figure 11: Absolute error of estimates (relative to Monte Carlo estimate of ground truth truncated mean absolute error) for the holdout (blue), 1NN (green) and our SNN (orange). We see that the holdout has significantly higher error in estimating the test risk in this task, which is caused by bias in the estimate provided.

E.5.3. Estimation of Ground Truth

We next describe why we might expect empirical test risk to provide a reasonable estimate of ground truth in this problem, and in particular we discuss the assumptions that justify the use of empirical test risk in relation to App. E.1. We consider a truncated loss $\ell(a, b) = \min(0, 10^6)$, and so Z_m is almost surely bounded by 10^6. The assumption that the $(Z_m)_{m=1}^{M^{\text{test}}}$ are independent would be implied by independence of $(Y_m)_{m=1}^{M^{\text{test}}}$ that is: that given the location at which a flat is sold, any remaining randomness in the observed sale prices is independent. Concretely, we might think of the randomness in sales price, ϵ_m^{test} in our model, as coming from aspects of the sale process of the house, such as who happens to see the advertisement for a house, and we *assume* these are independent for each house when constructing our estimate of the ground truth.

If the conditional independence assumption proposed above holds, then following App. E.1 we have that

$$|\hat{R}_{Q^{\text{test}}}(h) - R_{Q^{\text{test}}}(h)| \leq £1000000\sqrt{\frac{1}{2 \times 1000} \log \frac{2}{0.05}} < £43000. \tag{137}$$

Therefore, under this assumption, we might would expect our estimate of ground truth to be at least good enough to tell the difference (in the sense of which is closer to the actual ground truth) between predictors that differ in absolute error from the ground truth estimate by more than $2 \times £43,000 = £86,000$. Fig. 11 shows the absolute error of the three methods. We see that the error in the estimate provided by the holdout is on the order of $£150,000$, while the error in 1NN and SNN are closer to $£25000$ in most seeds. Given this large difference and earlier discussion, we do not expect that this error arises from difficulties in estimating the ground truth, but instead arises from actual differences in the qualities of the estimator.

E.5.4. Risk Estimation Details

Holdout, 1NN and SNN are run as in the air temperature experiments (App. E.4.5). In particular, we use a failure probability of $\delta = 0.1$ for SNN and nearest neighbor calculations are done with respect to Haversine distance to account for Earth's curvature. We use a fixed δ since we only discuss a method applicable to estimating fill distance on the unit cube, not on a subset of the sphere. $\Delta = £1,000,000$ is used in selecting the number of neighbors as this is an upper bound on the truncated loss. We use a Lipschitz constant of £1,000/km as £1/km seems implausibly small.

E.6. Wind Speed Prediction Experiment

In this section, we provide additional details for the wind speed prediction experiment discussed in Section 6.4 of the main text.

E.6.1. Data Sources and Pre-processing

We download daily wind speed readings from weather stations from the Global Historical Climate Network (Menne et al., 2012). As this dataset was constructed by NOAA employees, we understand it to be public domain following section 105 of the Copyright Act of 1976. We filter out weather stations outside the continental US, as well as any weather stations that do not contain daily average wind speed readings. We look only at wind speed data from January in the prediction task, and years 2000–2024. There is a weather station at Chicago O'Hare which we remove from the training and validation data and use as the test set.

For each replicate used to form Fig. 10 we split off a training set containing 80% of weather stations, and a validation set containing the remaining 20%. The number of observations in the training and validation set varies (because different weather stations may be online for a different number of days in January in previous years), but this leads to on the order of 580000 training observation and 126000 validation observations. Each training and validation point is a triple containing latitude, longitude and average wind speed. We perturb the latitude and longitude (in degrees) of validation points by a Gaussian random variable with standard deviation 10^{-12}, which is essentially equivalent to using random tie-breaking in the nearest neighbor algorithms. We expect this has a significant impact on 1NN compared to the version discussed in the paper, because there may be many observations from the nearest weather station to Chicago O'Hare. While this would unlikely be done in practice, random tie-breaking, or tie-breaking by selecting the first nearest neighbor according to some other ordering are common and would lead to similar outcomes as the results presented here (but higher variance than averaging over all neighbors that are equally close). The latitude and longitude are then converted to radians for the analysis.

E.6.2. Loss Function

We use truncated mean squared error as the loss function, $\ell(a, b) = \min(25, (a - b)^2)$.

E.6.3. Estimation of Ground Truth

Following the argument in App. E.1, we use the empirical test risk Eqn. (114) in place of the test risk as ground truth in synthetic experiments. The assumption that $(Z_m)_{m=1}^{M^{\text{test}}}$ are independent. It is sufficient for Y_m^{test} (the wind daily wind speeds) to be independent, conditioned on the spatial location at which it is observed. This is likely not the case, as we would expect average wind speed in consecutive days exhibit at least some dependence. However, if the wind speed decorrelates reasonably rapidly over time, we would expect similar arguments to hold, possibly with fewer effective samples.

Because the loss is truncated mean squared error, the Z_m are surely bounded by $25\text{m}^2/\text{s}^2$. We use 775 samples in estimating the risk. Using these numbers, and under the assumption that wind speed at a location is independent of the wind speed on previous days, in Eqn. (117), we arrive at

$$|\hat{R}_{Q^{\text{test}}}(h) - R_{Q^{\text{test}}}(h)| \leq 25\text{m}^2/\text{s}^2 \sqrt{\frac{1}{2 \times 775} \log \frac{2}{0.05}} < 1.22\text{m}^2/\text{s}^2. \tag{138}$$

Comparing to Fig. 12, we see that this application of Hoeffding's inequality is not sufficient to justify that the estimate of ground truth is accurate enough to attributed the observed better performance of SNN to (actually) better estimation of the ground truth as opposed to inaccuracies of our Monte Carlo estimate of the test risk. However, we expect this is largely due to looseness is Hoeffding's inequality and, given that a large difference is observed in most seeds, it would be very surprising if this was only due to error in estimation of the ground truth which is independent across seeds.

E.6.4. Model Fitting

A gradient boosted machine is fit using LightGBM (Ke et al., 2017) with default parameters expect for the number of leaves (set to 127) and the number of estimators (set to 100).

E.6.5. Risk Estimation procedures

Holdout, 1NN and SNN are run as in the air temperature experiments (App. E.4.5). In particular, we use a failure probability of $\delta = 0.1$ for SNN and nearest neighbor calculations are done with respect to Haversine distance to account for Earth's curvature. We use a fixed δ since we only discuss a method applicable to estimating fill distance on the unit cube, not on a subset of the sphere. A Lipschitz constant of $1(\text{m}^2/\text{s}^2)/\text{km}$ is used for selecting the number of neighbors.

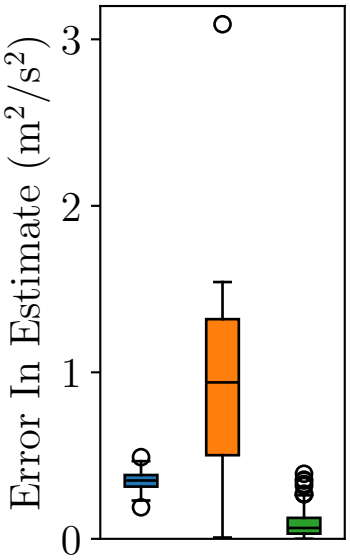

Figure 12: Absolute error in estimating (approximate) test risk in the wind speed experiment for the holdout (blue), 1NN (orange) and our SNN (green). It appears that SNN has the smallest error in estimating the ground truth, although the approximate ground truth we compute via Monte Carlo estimate is not entirely theoretically justified. However, due to the relatively large differences across most seeds, we still expect the difference is indicative of better performance of SNN.

E.7. Model Selection on Synthetic Data

We next see that SNN and 1NN are able to select the model with lower test risk in a model selection task, but the holdout systematically picks the wrong model. We repeat the model selection problem 100 times. In each repetition, we have $N^{\text{train}} = 100$ and a max $N^{\text{val}} = 75$. In our analysis, we will consider validation subsets of size $N^{\text{val}} \in \{5\ell\}_{\ell=1}^{15}$. We generate independent test, validation, and training data as follows; see Fig. 13.

$$\begin{aligned} U_i^j &\sim \mathcal{U}([-0.5, 0.5]), \quad S_i^j = \sqrt{U_i^j + 0.5}, \quad j \in \{\text{train}, \text{val}\} \\ S_m^{\text{test}} &= m/20 - 0.5, \quad 0 \le m \le 20, \quad \epsilon_i^j \sim \mathcal{U}([0, 0.1]), \\ Y_i^j &= |S_i^j| + \epsilon_i^j \quad j \in \{\text{train}, \text{val}, \text{test}\}, \end{aligned} \tag{139}$$

We compare two predictive methods: $h_0(S) = 0.25$ and $h_1(S) = \beta_1^\top S + \beta_0$, with (β_1, β_0) fit by minimizing the mean absolute residual on the training data. Fig. 13 shows the data and predictions of both models (as functions of space). We use the loss function $\ell(a, b) = |a - b|$, which is bounded for this problem because both the hypotheses and the response variable are bounded on $[0, 1]$.

Across all seeds, h_0 has the lower empirical test risk; h_1 makes large errors on the test points near 0 because most of the training data is in $[0, 0.5]$. Since most of the validation data also clusters near 1, we expect the holdout to struggle due to bias. Our arguments in App. C lead us to expect both SNN and 1NN should perform well on this task when given sufficient validation data.

We say an estimator of the risk, $\hat{R}$, selects h_0 if $\hat{R}(h_0) < \hat{R}(h_1)$. We plot the percentage of times each method (correctly) selects h_0 as a function of the number of validation points in Fig. 14. When the validation set is small, all estimators select the model with lowest test risk (h_0) less than half the time. For the nearest neighbor methods, we expect that when there are few or no spatial locations less than 1, weighting cannot fix the estimate. However, when the number of validation points is large, the nearest neighbor methods consistently (correctly) select h_0. By contrast, the holdout consistently (incorrectly) selects h_1, even though h_1 has higher test risk. See App. C for full experiment details.

Data Generation The data generation is fully described by Eqn. (139).

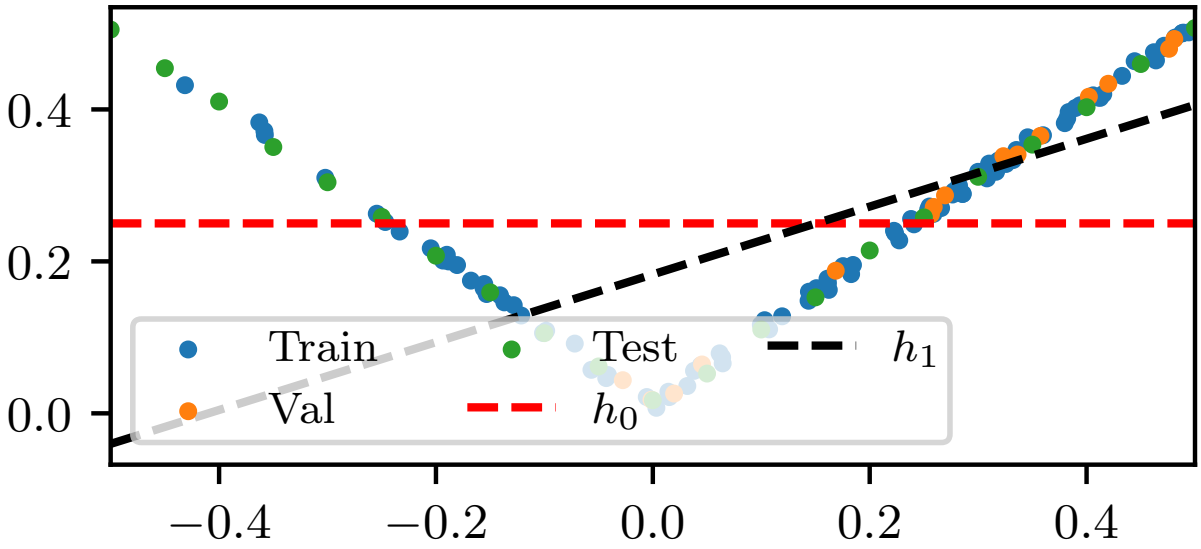

Figure 13: Training (blue), validation (orange), and test (green) data for a single seed of the model selection experiment. The dashed red line depicts predictive method h_0, and dashed black shows h_1.

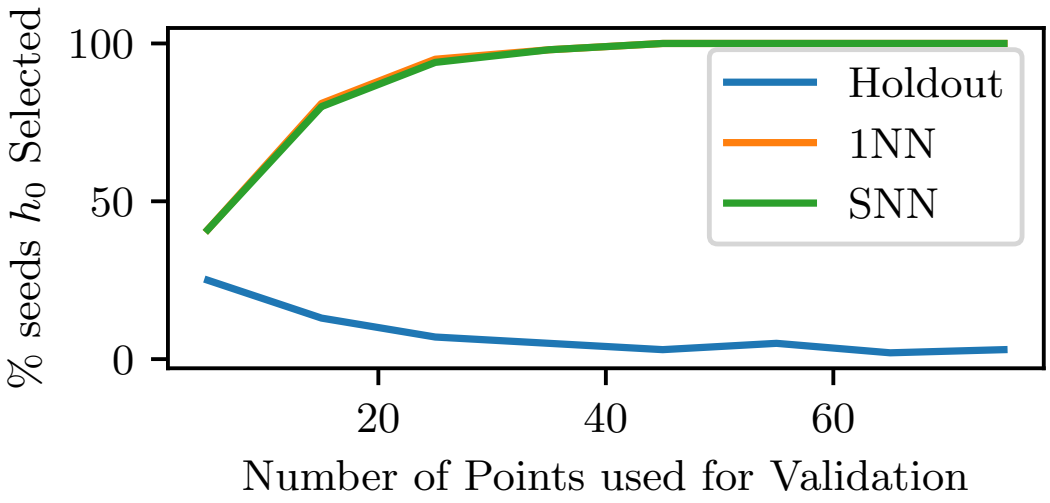

Figure 14: The percentage of times each estimator (correctly) selects the model with lower empirical test risk (h_0), out of 100 seeds as a function of N^{val}. Estimators include the holdout (blue), 1NN (orange), and our SNN (green).

Model Fitting We consider two models. The first is a constant predictor that predicts 0.25. The second is an affine model (a linear model with an intercept) fit by minimizing the mean absolute error from the line to the training points. This is fit using the `Scikit-learn` quantile regression with the (default) "HiGHS" solver (Huangfu & Hall, 2015).

Estimation of Risk The validation estimates used are calculated in the same as the synthetic experiments outlined previously in App. E.3.4. We use $\Delta = 1$ in the bound when selecting $k^\star_{T_2}$, even though the absolute value loss used can be larger than 1. We don't expect this to have a significant impact on the results, as the upper bound we minimize is already misspecified in a similar way by not using the actual Lipschitz constant of the function. We again use $\delta = 0.1$ when selecting $k^\star_{T_2}$.